ONLINE PROCESSING OF SCALAR IMPLICATURES IN CHINESE AS REVEALED BY EVENT-RELATED POTENTIALS

BY

Stephen Politzer-Ahles

Submitted to the graduate degree program in Linguistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts

Committee:

_____

Chair, Robert Fiorentino

_____

Alison Gabriele

_____

Jie Zhang

Date defended: 11 April 2011

The Thesis Committee for Stephen Politzer-Ahles

certifies that this is the approved version of the following thesis:

ONLINE PROCESSING OF SCALAR IMPLICATURES IN CHINESE AS REVEALED BY EVENT-RELATED POTENTIALS

_____

Chair: Robert Fiorentino

_____

Alison Gabriele

_____

Jie Zhang

Date approved: 25 April 2011

Abstract

During sentence processing, whether pragmatic information is integrated immediately and automatically or at a delay is a subject of debate in experimental pragmatics. One test case is that of scalar implicatures, which occur in statements like "some of the students are hardworking", which have both a *logical* meaning ("at least one is hardworking") and a *pragmatic* meaning ("not all of them are hardworking"). Default processing accounts hold that the pragmatic meaning of *some* comes online immediately and effortlessly, whereas context-based processing accounts propose that this meaning is not generated until after the logical meaning.

Previous event-related potential (ERP) studies on scalar implicatures typically investigated critical words downstream of the quantifier and were thus not able to address the possibility of immediate construction of scalar interpretations at the moment the quantifier is encountered. Furthermore, effects of lexico-semantic processing and real-world context make it difficult to interpret effects observed in these studies. The present study adopts a picture-sentence design to make the violation immediately detectable when the quantifier is read and to control the context in which the sentence is understood. Participants saw pictures in which several characters are either performing the same activity or different activities, followed by sentences using “some” or “all”, yielding a 2x2 design including both pragmatic violations (“some” sentences after “all” pictures) with matched controls, and purely incorrect assertions (“all” sentences after “some” pictures) with matched controls. Crucially, the pragmatic violation cannot be recognized as a violation until after the pragmatic meaning of *some* is computed.

Pragmatic violations and purely logic violations elicited an early N400 effect and a right-lateralized negativity in the 600-900 ms time window, whereas purely logic violations elicited qualitatively different effects in at least the late time window. These results demonstrate that the

pragmatic meaning of *some*, which relies on the generation of a scalar implicature, is available to the processor immediately; furthermore, they show that errors based on pragmatic expectations and errors based purely on logic elicit qualitatively different electrophysiological responses. I conclude that these findings are consistent with a default processing account, although they do not rule out a context-driven account.

Acknowledgements

I would like to thank my advisor, Robert Fiorentino, for his guidance in all aspects of experiment design, data collection, data processing and analysis, and data reporting for this experiment, as well as for being the one "activate" my interest in neurolinguistics in the first place.

I would also like to thank Jiang Xiaoming for countless hours spent discussing the design of this experiment with me; the concept for the present work is as much his idea as mine. The preparation and stimulus design for the present study, as well as data collection for later studies, was conducted at the Center for Brain and Cognitive Sciences at Peking University under Zhou Xiaolin, and I am grateful for his support and permission to use his facilities and participate in his laboratory. This research was funded by the NSF's East Asia and Pacific Summer Institutes program, without which I would not have had the opportunity to collaborate with this laboratory.

I owe thanks to numerous colleagues and classmates for their help in stimulus design and testing: Liang Yan, Wu Chunping, and Wu Yue provided the vast majority of the materials used in this experiment, and Jiang Liu, Hongying Xu, Yuanliang Meng, Yingjie Li, Grace Zhang, Huang Xiaoli, and Ji Xianghan put up with frequent solicitations for acceptability judgments as the stimuli were being edited.

I would not have been able to complete the data collection, analysis, and interpretation of the data for this study without the help of my classmates in the University of Kansas linguistics department, particularly my labmates in the Neurolinguistics and Language Processing Laboratory (thanks to Hyunjung Lee and Lamar Hunt for help with data collection, and to José Alemán Bañón, Kristi Bond, Yuka Naito-Billen, Jamie Bost, and Ella Fund-Reznicek for showing me the ropes of data collection), and my classmates in Rob Fiorentino, Alison Gabriele, and Utako Minai's co-taught Research in Acquisition and Processing seminar. I am also grateful to Chloe Shen who, through QQ wizardry, single-handedly recruited over half of my participants. Finally, thanks to Wu Yin, Zhou Yuqin, and Zhang Huihui for helping collect the data that is reported in Appendix I.

Table of Contents

List of figures

List of tables

1 Introduction

When engaging in natural speech we must process not only the syntactic and semantic content of the other speaker's input, but also compute how his or her utterances relate to the context in which they are being made, and to our own expectations about what should be said. These types of information fall under the broad heading of pragmatics. Compared to the processing of syntactic and semantic information, the processing of pragmatic information in the brain is not well understood. The emerging fields of experimental pragmatics (Noveck & Reboul, 2008) and neuropragmatics (Van Berkum, 2010) seek to better understand how we use pragmatic information to understand language and whether the processing of such information is distinct from the processing of, e.g., syntactic and semantic structure. In particular, the pragmatic case of *scalar implicature* has received attention in numerous recent experimental studies (Katsos & Cummins, 2010; Tavano, 2010; Noveck & Reboul, 2008; Noveck & Sperber, 2007). This thesis presents an electrophysiological experiment conducted to elucidate how scalar implicature is processed in real time.

1.1 Overview of scalar implicature

Scalar implicature refers to the interpretation of terms which fall on the low end of a semantic "scale" as implying the negation of terms which fall higher on the scale. Terms including, but not limited to, quantifiers, adjectives, numbers, modals, and conjunctions may be scalar (Katsos & Cummins, 2010; Rullman & You, 2006). For example, consider the following dialogue:

1) A: Are the students in this department hardworking?
   B: Some of them are.

In B's statement, the implication is that *not all* of the students are hardworking. The *not all* meaning, however, is not part of the inherent semantics of the quantifier *some*; rather, it is the "pragmatic meaning" of the quantifier, as opposed to its inherent meaning (the "logical" or "semantic") meaning, which is simply *at least one*. The fact that the *not all* meaning comes about through pragmatic implicature and is not part of the quantifier's logical meaning can be demonstrated by the fact that it can be undone without resulting in illogicality (i.e., the implicature is defeasible):

2) Some of the students in this department are hardworking. In fact, all of them are.

On the other hand, the logical meaning of *some* is not based on a pragmatic implicature and thus is not defeasible, as shown in (3):

3) Some of the students in this department are hardworking. #In fact, none of them are.

If *not all* were part of the inherent meaning of "some", then (2) would be nonsensical like (3) is.

Some is also defeasible in prediction contexts. Consider the following examples:

4) Speaker A: If you find some of the Easter eggs, you'll win a prize.
Speaker B: I found all of them!

In example (4), speaker B has met the conditions of the game and will win a prize. This example again demonstrates that *not all* is not an inherent part of the meaning of *some*, because *all* can still satisfy the truth conditions of a *some* proposition.

Because the *not all* reading of *some* does not appear to be part of the word's inherent semantics, it is assumed to be derived from a pragmatic inference. Specifically, following the theory proposed by Horn (1972), the quantifier "some" is thought to exist on a scale in which it is weaker (less informative) than the quantifier "all", which is higher on the scale and therefore stronger (more informative). If we believe that the speaker is fully knowledgeable of the

situation, cooperative, and willing to use the most informative term possible (Grice, 1975), then when the speaker says "some" we infer that the speaker was unable to say "all", and thus that "all" must not be true (Katsos & Cummins, 2010; Nieuwland et al., 2010; Wu & Tan, 2009; Noveck & Sperber, 2007; Noveck & Posada, 2003). Thus, a quantifier such as "some" is thought to have both a *logical* meaning ("at least one") which is inherent to the quantifier's semantics, and a *pragmatic* meaning ("at least one, and not all"), which is derived through a pragmatic computation taking into account the context, and the intentions and abilities of the speaker (Katsos & Cummins, 2010; Wu & Tan, 2009, Noveck & Posada, 2003).

The process by which speakers derive this inference is the focus of most of the experimental studies on scalar implicature. Broadly speaking, *default models* hold that scalar implicatures for terms like "some" are generated automatically, without the need to consider the context and the speaker (but may be cancelled later if a special context indicates that the logical meaning of the term is the relevant one, as in example (2)); whereas *context-driven models* hold that scalar implicatures are only generated when the context calls for it, and thus that they are delayed until the context can be processed (for reviews, see Katsos & Cummins, 2010; Tavano, 2010; Noveck & Sperber, 2007).

On-line psycholinguistic studies have commonly investigated reading time slowdowns or other processing costs associated with generating (and, if necessary, cancelling) scalar implicatures (e.g. Breheny et al., 2006; Tavano, 2010), the relative latency of behavioral measures associated with processing the pragmatic and logical interpretations of scalar terms (e.g. Grodner et al., 2010; Degen & Tanenhaus, 2010b; Degen, 2009; Huang & Snedeker, 2009; Bott & Noveck, 2004; Noveck & Posada, 2003), and contextual conditions influencing whether or how quickly an implicature is generated (e.g. Bergen & Grodner, 2010; Breheny et al., 2010).

The results of these studies are equivocal; while the majority have shown reading time or eye movement delays and argued that the pragmatic meaning of *some* becomes available later than the logical meaning and that generating a scalar implicature engenders a processing cost (e.g. Noveck & Posada, 2003; Breheny et al., 2006; Huang & Snedeker, 2009; Degen, 2009; Tavano, 2010), others have shown that, at least under certain conditions, the pragmatic reading of *some* seems to be processed as quickly and effortlessly as the logical meaning (e.g. Grodner et al., 2010; Degen & Tanenhaus, 2010b). But behavioral results regarding scalar implicature generation, particularly response times in verification tasks and reading times in self-paced reading, are difficult to interpret given that they reflect not only processing related to implicature generation but also controlled decision-making components (Nieuwland et al., 2010; Tavano, 2010). Thus, it is worthwhile to investigate these questions using a methodology that provides more fine-grained temporal resolution. One such methodology is eye-tracking, which has been adopted in many recent studies (Tavano, 2010; Grodner et al., 2010; Degen & Tanenhaus, 2010; Huang et al., 2010; Huang & Snedeker, 2009; Degen, 2009). Another is electroencephalography, which is the method used in the present study. In what follows, I will first briefly review several relevant behavioral and eye-tracking findings regarding the processing of scalar implicatures, and will then review in more depth the extant electrophysiological findings.

1.2 Behavioral and eye-tracking studies of scalar implicature processing

The critical questions addressed in previous studies on scalar implicature are a) how quickly the "pragmatic reading" of a scalar term (e.g., *not all* for the quantifier *some*) becomes available to the parser, and b) whether realizing the pragmatic reading engenders a processing cost. Breheny and colleagues (2006, Experiment 3) investigated the second of these questions in

a self-paced reading study comparing quantifiers in upper-bounded contexts—in which the pragmatic meaning of the quantifier is relevant and thus expected to be generated, as in (5a)—against those in lower-bounded contexts—in which the pragmatic meaning is not relevant and thus not expected to be generated, as in (5b):

5a) Mary asked John whether he intended to host all of his relatives in his tiny apartment. John replied that he intended to host some of his relatives. (The rest would stay in a hotel.)

5b) Mary asked John why he was cleaning his apartment. John replied that he intended to host some of his relatives. (The rest would stay in a hotel.)

In their study, the critical phrase "some of his relatives" elicited longer reading times in upper-bounded than lower-bounded contexts; the authors took this result to indicate that generating the pragmatic meaning of the quantifier engenders a processing cost, and thus that scalar implicatures are not generated by default. The stimuli, however contain several other confounding factors that may account for this finding; furthermore, the nature of the self-paced reading task makes it difficult to determine what specific processing demands lead to the increased reading time (see Huang & Snedeker, 2009, p. 381, for a review of these points). Thus, a number of later studies have adopted the visual world eye-tracking method, which provides more fine-grained temporal data.

Huang and Snedeker (2009) used a visual world design that allowed the investigators to see how early the pragmatic reading of a quantifier becomes available by measuring when participants begin to fixate on an item in a visual array. They showed participants visual arrays which included, among other things, a character who has all of an item (e.g., a girl holding all the soccer balls in the array) and another character who has only some of another item (e.g., a girl holding some of the soccer balls in the array, while the rest are held by a boy); see Figure 1 for an example. While participants inspected the arrays, they were presented an auditory sentence

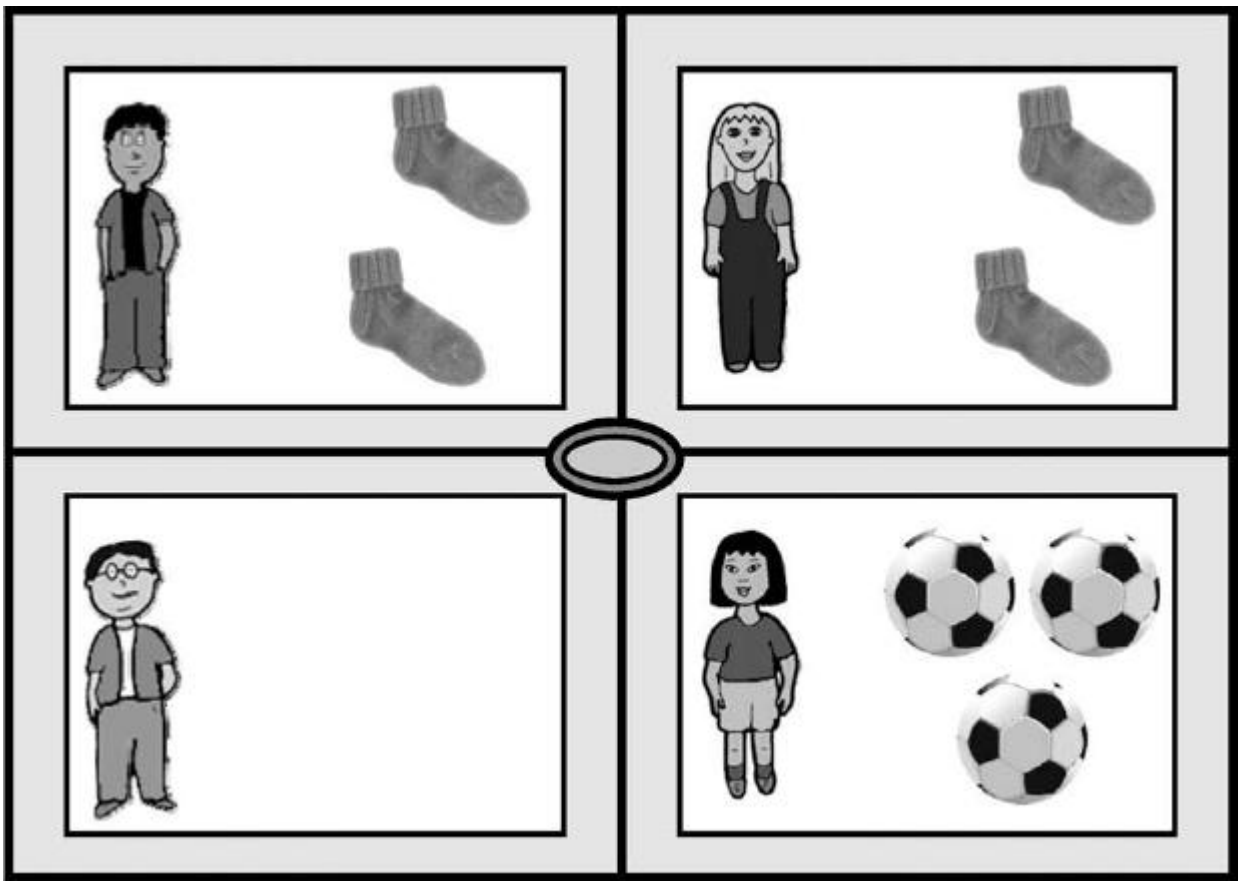


Figure 1: Sample visual array used in Huang & Snedeker (2009)

such as, "Point to the girl who has some of the socks". When the quantifier is heard, both the girl with all the soccer balls and the girl with only some of the socks are possible candidates; if the pragmatic meaning of *some* is computed immediately, however, the participants ought nonetheless to quickly fixate on the girl with only some of the socks, since "some of" is not pragmatically consistent with the picture of the girl holding all of the balls. Huang and Snedeker, however, found that participants looked to each of these pictures an equal proportion of the time, and did not start looking mainly at the picture of the girl holding some of the socks until after the disambiguating noun "socks" was heard. On the other hand, when the sentence "Point to the girl who has all of the soccer balls" was played, as soon as the quantifier was heard participants

fixated immediately on the appropriate picture. The authors take this result as support for the notion that pragmatic meanings are accessed at a delay relative to logical meanings.

Grodner and colleagues (2010) report a study using essentially the same design as Huang and Snedeker (2009) but with several changes made to the stimuli. First of all, Grodner and colleagues (2010) present sentences with a phonetically shortened quantifier *summa,* rather than *some of,* based on Degen and Tanenhaus' (2010a) finding that the former is more likely to elicit a scalar interpretation, whereas the latter may be interpreted as an existential until the partitive *of* is heard. Secondly, the stimuli were adjusted to make the scalar implicature more relevant and more salient, particularly by not including items with other quantifiers such as numbers (as in "the girl who has two of the socks"; see Degen and Tanenhaus, 2010a, for discussion of how the alternatives available in an experimental context affect the interpretation of *some*) and by prefacing each experimental sentence with an auditorily-presented description of the scene. Thirdly, the researches included several control conditions that were not included in Huang and Snedeker's (2009) design. The results of Grodner and colleagues' (2010) study were the opposite of those of Huang and Snedeker's (2009): namely, looks to the target increased immediately when *summa* was heard, and were not delayed relative to looks to target in sentences that did not require computing scalar implicature. Because of the large number of differences between the materials for the two studies, it is difficult to identify a single factor that may have caused the difference, but the authors suggest that the results at least demonstrate that there are conditions under which the pragmatic meaning of *some* emerges immediately (as in their study) and conditions under which it does not (as in Huang and Snedeker's (2009)).

Further research has suggested that the presence of alternative quantifiers in the experimental context may influence the speed at which scalar implicature on *some* is processed.

A study in German by Degen (2009) using a mostly similar paradigm but manipulating properties of the stimuli found that awareness of the pragmatic reading of (German) *some,* indicated by proportion of fixations on the target, emerged later than awareness of the inherent meaning of (German) *all* but earlier than the noun itself, suggesting that implicature was generated at a delay but still generated online quickly enough to provide disambiguation in the proper context. Crucially, this study included filler trials using numbers rather than quantifiers, decreasing the consistency of the mapping between images and quantifiers and making "some" less predictable. Another study by Huang and colleagues (2010) also reduced the predictability of "some" in the same way and again found that the pragmatic meaning was generated at a delay, further supporting the notion that experimental contexts plays a substantial role in how quickly scalar implicatures are generated. When they specifically manipulated predictability in a between-subjects design, though, they found that the implicature was realized at a delay in both conditions, which suggests that the delay in implicature processing may be due to other factors (if predictability alone were the key factor, one would expect the pragmatic reading of the quantifier to come online quickly in the high-predictability condition and slowly in the low-predictability condition). Thus, while several authors have asserted that this feature of experimental context should influence the speed at which implicatures are computed (Grodner et al., 2010; Huang et al, 2010), it remains to be confirmed experimentally.

Tavano (2010) used another type of eye-tracking design to investigate a variety of eye-movement measures that may elucidate the time course of scalar implicature generation and processing costs associated with it. She showed participants visual arrays in which either all of the items shared the same property (left side of Figure 2) or ones in which the items differed on that property (right side of Figure 2). She paired these pictures with auditory sentences

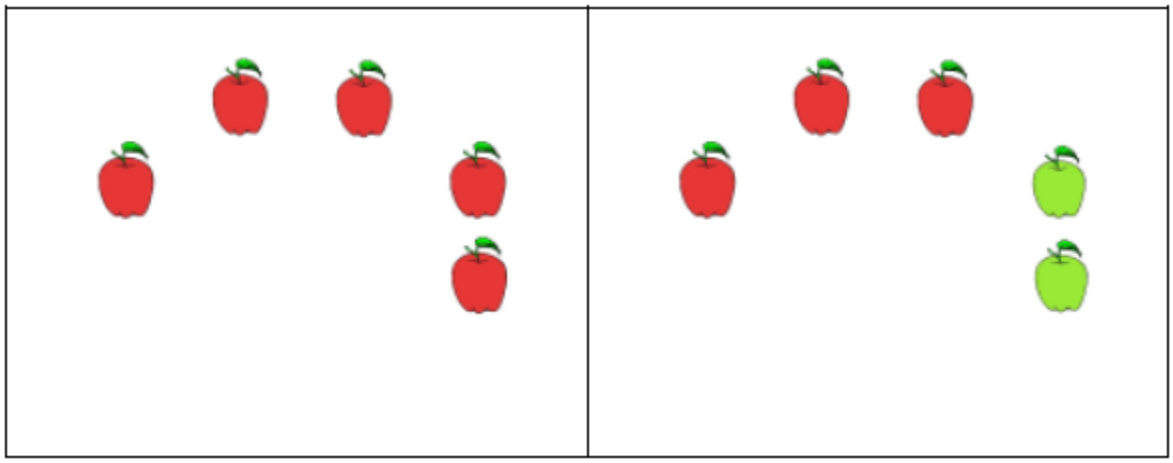


Figure 2: Sample visual arrays used in Tavano (2010); the left array is typical of an "all"-type picture, the right typical of a "some"-type picture.

describing the array using either *some* (e.g., "This is a picture of apples; some of them are red") or *all* ("This is a picture of apples; all of them are red") and recorded participants' eye movements as they listened to the sentences and inspected the arrays. Tavano found that participants become aware of the pragmatic interpretation of *some* early, given that when they were inspecting a "some"-type picture and heard a "some"-type sentence, they began inspecting the other group of objects (e.g., if they had been fixating on the green apples first, they looked to the red apples once they heard *some*) earlier than if they heard an "all"-type sentence, suggesting that the quantifier *some* made them quickly pay attention to differences in the set of items. On the other hand, processing the scalar quantifier also engendered a processing cost, as evidenced by longer inspection times when hearing *some* than when hearing *all*. Tavano thus argues that the results do not completely support either class of psychopragmatic models: default accounts do not predict that implicature generation would engender a processing cost, whereas context-driven accounts do not predict that the implicature would be generated in this task (where, she argues, the implicature is not relevant).

It can be seen, then, that behavioral and eye-tracking results so far are equivocal with regard to the speed of scalar implicature generation. The present study adopts another methodology, event-related potentials, which I review in the next section.

1.3 ERP studies of scalar implicature processing

The event-related potential (ERP) technique is a promising means to investigate the time course of scalar implicature generation. This technique measures changes in scalp-recorded electroencephalogram (EEG) voltage that are time-locked to the presentation of a stimulus (Coulson, 2007). ERPs, like eye-tracking, offer a dynamic view of how processing unfolds through time and how the EEG changes from one millisecond to the next; therefore, ERPs offer a more fine-grained view of the time course of the language processing than offline judgments, reaction times, and even self-paced reading times, which generally only provide information about the final output of some process (Bornkessel-Schlesewsky & Schlesewsky, 2009; Luck, 2004). ERP experiments on sentence processing often use a *violation design* (Bornkessel-Schlesewsky & Schlesewsky, 2009), in which participants are presented with stimuli that violate some linguistic rule or expectation (e.g., ungrammatical or semantically anomalous sentences); the logic is that if a violation of a certain type of linguistic information elicits a certain ERP response, the latency of that response provides an upper bound on when that type of information must have been processed. Additional features of the ERP, such as its polarity and scalp topography, allow the researcher to speculate about whether that information was processed in the same way or by the same neuronal populations as other types of linguistic information.

Two previous ERP studies have analyzed the electrophysiological responses to sentences that included as the violation an infelicitous instance of scalar implicature, in which a sentence using "some" was logically correct but pragmatically underinformative—that is to say, given the context, the logical reading of the sentence was true but the pragmatic reading (the reading in which *some* is interpreted as meaning *not all*) was not. Noveck and Posada (2003) measured ERPs elicited by the sentence-final words in the French equivalents of patently true, patently false, or underinformative sentences, as exemplified in (5):

5) a. Some gardens have trees.
   b. *Some toads have churches.
   c. #Some dogs have ears.[1]

The underinformative sentences (5c) were correct under a logical interpretation (there do exist dogs that have ears) but incorrect under a pragmatic interpretation (it is not the case that "not all dogs have ears"). The investigators found that, although some participants generally responded pragmatically and some generally responded logically, the logical responders had faster reaction times in all conditions; the investigators took this to mean that scalar implicatures take time and effort to calculate. In the ERP data, for both groups of participants they found a decreased N400 component for underinformative sentences relative to other conditions, and no difference in ERP responses between the true and false conditions. The N400, a negativity with a broad or slightly right-lateralized distribution over centro-parietal recording sites, usually with an onset around 250 ms and a peak around 400 ms after stimulus presentation, is typically associated with semantic processing or integration, and is more negative when a word is unexpected, semantically anomalous, or otherwise difficult to retrieve from the lexicon or integrate with

[1] In this paper, I indicate false or semantically anomalous utterances with an asterisk (*) and pragmatically infelicitous ones with a hash (#). Words to which ERPs were time-locked are underlined.

context (Lau et al., 2008; Kutas & Federmeier, 2000). The authors interpret this finding as suggesting that underinformative sentences elicit decreased semantic integration, but this finding is difficult to interpret for several methodological reasons, including differences in lexico-semantic relatedness between subjects and objects in their materials, the lack of counterbalancing or controlling between critical items, the fast stimulus presentation rate, and the possible effect of sentential wrap-up processes (for a review of these concerns, see Nieuwland et al., 2010; for a discussion of sentence wrap-up processes, see Hagoort et al., 2003). It is also concerning that semantically anomalous sentence-final words in the patently false condition (5b) did not elicit an N400 effect, which would normally be expected (Nieuwland et al., 2010).

A later study by Nieuwland and colleagues (2010) addressed these methodological factors and additionally compared responses between participants group with high and low pragmatic abilities, and conducted Latent Semantic Analysis (Landauer et al., 1998) on the critical words in the informative and underinformative conditions to test whether the obtained results were modulated by semantic relatedness. The latent semantic analysis (LSA) showed that the informative critical words, relative to the underinformative ones, were usually less semantically related to the subjects of their respective sentences. Experimental materials are exemplified in (6):

6) a. Informative, less related: Some people have pets, which require good care.
   b. Underinformative, more related: #Some people have lungs, which require good care.

Furthermore, they examined violations of pure lexico-semantic fit by comparing high vs. low cloze probability words, as in the following pair:

7) a. High cloze: Wine and spirits contain alcohol in different amounts.
   b. Low cloze: Wine and spirits contain sugar in different amounts.

Whereas Noveck and Posada (2003) asked participants to make explicit judgments about the truth-value of the stimulus sentences, Nieuwland and colleagues' (2010) participants read sentences passively. In the ERP results, they found that participants with high pragmatic abilities showed an increased N400 effect in response to the underinformative terms, relative to the informative ones, whereas participants with low pragmatic abilities showed the opposite effect (an increased N400 in response to informative terms). The N400 effect was modulated by LSA relatedness for the low-ability group, but not the high-ability group, such that participants with low pragmatic ability only had an N400 effect when the informative term was significantly less semantically related to the sentence subject than the underinformative turn. Taken together, these results suggest that the high-ability participants were sensitive to the scalar implicature, whereas the low-ability participants were only sensitive to lexico-semantic relatedness (which is to say, words related to the subject were primed by the subject, which for low-ability participants reduced the N400 amplitude regardless of sentence context; for further discussion of the differential effects of priming and sentence context on the N400, see Urbach & Kutas, 2010; Nieuwland & Kuperberg, 2008; Fischler et al., 1983; Kutas & Hillyard, 1984). On the other hand, the critical words in the lexico-semantic fit manipulation elicited a classic N400 effect for both participant groups regardless of pragmatic ability, suggesting a dissociation between pragmatic and semantic processing. In a second experiment (Nieuwland et al., 2010, Experiment 2), the authors found that the pragmatic N400 effect disappeared when the critical word was taken out of discourse focus by being followed by a restricting relative clause that made the sentences globally informative (e.g., *Some people have pets/lungs that are diseased by viruses.*). Although the participant seeing the critical word could not yet know that the word was going to be de-focused by a following relative clause, the authors argue that because all the critical

sentences in Experiment 2 followed this design, an experimental context was created in which participants were discouraged from generating scalar implicatures. The results of these two experiments suggested that scalar implicatures can be processed immediately—a result consistent with findings in visual world eye-tracking (Grodner et al., 2010; Tavano, 2010; Degen & Tanenhaus, 2010b; but see Huang & Snedeker, 2009; Huang et al., 2010)—and can guide expectations about upcoming linguistic input, but that scalar implicatures are apparently not generated when context cues indicate that they will be irrelevant to the interpretation of the sentence.

Nieuwland and colleagues (2010) also performed an exploratory analysis of ERP responses to quantifiers, in which they compared ERPs elicited by *some* to those elicited by *many* (in filler sentences), which they argue may elicit a weaker *not all* interpretation than *some* does. They found that, for participants with low pragmatic ability, *many* had a slightly increased negativity relative to *some* on right-lateralized recording sites in the 300-350 ms time window, and that it had a decreased positivity relative to "many" on frontal recording sides in the 650-700 ms time window; in their second experiment, the same comparison yielded a slightly increased negativity for "some" on anterior electrodes in the 450-500 ms time window. They argue that the participants with low pragmatic ability may have failed to show an N400 effect for underinformativeness because they strategically suppressed the implicature, and that the presence of differential effects between *some* and *many* for these participants is consistent with such an account.

While these studies have pioneered the job of "electrifying" scalar implicature research and have contributed greatly to our understanding of scalar implicature processing, several conceptual factors limit the extent to which they can inform us about how scalar implicatures are

generated in real time. In these studies, violations only became noticeable (and thus effects only became detectable) on words well downstream of the quantifier, presumably after the pragmatic meaning of the quantifier had already been computed.[2] This is in contrast to studies using the visual world eye-tracking method (Tavano, 2010; Grodner et al., 2010; Huang & Snedeker, 2009) and a study using self-paced reading of stories (Breheny et al., 2006), which have compared conditions at moment the quantifier is encountered. Some theories of scalar implicature hold that "some" is a is a term that triggers a Generalized Conversational Implicature (Grice, 1989), which means that its pragmatic meaning becomes available immediately and by default, rather than having to wait to see if the pragmatic meaning is relevant to the statement (for reviews see Katsos & Cummins, 2010; Noveck & Sperber, 2007). In that case, the N400 responses elicited in the ERP studies described above may not have indexed implicature generation *per se*, but rather the processing of unexpected words, where the expectation may have been created by a previously-computed scalar implicature; this is especially likely given that it has been shown that the N400 effect can be modulated by expectations based on negative-like quantifiers such as "few" (Urbach & Kutas, 2010).

An additional concern in using content words downstream of the quantifier has to do with lexico-semantic processing that is time-locked to these words but not related to scalar implicature processing. The results of the Latent Semantic Analysis performed by Nieuwland and colleagues (2010) revealed that the response to the processing of scalar implicature at the region tested in their study and in Noveck & Posada (2003) may be obscured by effects of

[2] Nieuwland and colleagues (2010) did conduct an exploratory analysis of the ERP responses time-locked to the presentation of the quantifier, which is discussed below; their experiment, however, was not designed to test this region, and thus it is not possible to compare responses to informative and underinformative sentences at the quantifier region in their data.

lexico-semantic processing of content words, which occurs in the same time window as the pragmatics-related response. Thus, it is worthwhile to apply to ERP technique to a design in which differential effects of generating a scalar implicature may be measured at the moment the quantifier, a non-content word, is encountered.[3]

Furthermore, the two previous ERP studies both used real-world knowledge as the context against which the truth, falsehood, or underinformativeness of propositions was evaluated. Thus, reading the critical words in these studies may have initiated not only lexical activation and comparison of the critical word against what was expected based on the context, but also a large-scale search through long-term memory to find members of the previously-stated category that do not match the quality given in the sentence. For instance, when seeing #*Some televisions have screens* (Noveck & Posada, 2003), a reader probably cannot judge the correctness or felicitousness of this sentence until he/she has searched through the category "televisions" to see if he/she can find any exemplars that do not have screens. Nieuwland and colleagues' (2010) stimuli were controlled across conditions for the size of the real-life category represented by the sentence subjects, since informative and underinformative stimuli were formed by manipulating the critical word within identical sentences. Noveck and Posada's (2003) stimuli, however, were not; in fact, some category size differences were present between conditions (e.g., *Some animals have stripes* in the patently true condition and #*Some giraffes have necks* in the underinformative condition), although these differences were not systematic. Furthermore, the use of real-world knowledge as a context introduces additional unpredictability

[3] Note that there is no principled reason that investigations of scalar implicature have to focus on non-content words like quantifiers and coordinators; many adjectives (e.g., <hot, warm>, <beautiful, pretty>) and verbs (e.g. <hate, dislike>, <love, like>) are also scalar (see Katsos & Cummins, 2010; Rullman & You, 2006). Quantifiers, coordinators, and numbers are simply the most common test cases adopted in the existing literature, and adopting quantifiers as a test case in the current study allows for the most straightforward comparison with previous studies.

and variability into participants' interpretations of the stimuli. In particular, some of the "underinformative" sentences used may not actually be underinformative for creative readers: for instance, after reading *#Some televisions have screens*, a reader may imagine a broken television whose screen has been smashed in; after reading *#Some cherries have pits*, a reader may imagine cherries that have already been pitted. It is difficult to tell if participants are doing such processing and how it may affect the obtained results. For these reasons, it would be ideal to strictly control the context in which experimental sentences carrying scalar terms are processed.

1.4 The present study

The present study adopts a picture-sentence verification design (Wu & Tan, 2009; Tavano, 2010) to address both of the abovementioned issues. In picture-sentence verification, participants read or hear a sentence after (Wu & Tan, 2009) or while (Tavano, 2010) viewing an image. The experimental materials may be manipulated such that the sentence correctly, incorrectly, or underinformatively describes the picture. Using such a design, it is possible both to strictly control the context in which the sentence is interpreted (and thus control the categories and objects necessary for participants to search through), and to make violations become detectable at the position of the quantifier itself. Like Wu and Tan (2009) and Tavano (2010), I create a 2x2 design by first presenting participants with a picture displaying several identical characters, in which either all of the characters are performing the same activity or some are performing one and the rest performing another ("all"-type or "some"-type pictures), and then presenting a sentence that either says "all of the X are Y" or "some of the X are Y" ("all"-type or "some"-type sentences). By the time the participant hears the quantifier "some", if all the activities shown in the picture are identical, the sentence is already pragmatically infelicitous, as

there is no sentence that can felicitously describe the picture using "some". When a participant hears "all", if the picture displays two different activities, the sentence can still be completed in a true way (e.g., if the picture shows some chefs chopping cucumbers and some shopping onions, a sentence beginning *All the chefs…* could still be completed as *All the chefs are wearing hats*); in the present study, however, no such sentences are concluded (the sentence always refers to the most salient part of the picture), and thus when a participant sees a "some"-type picture she or he ought to expect a "some"-type sentence. Thus, the present design allows a comparison between the effects of encountering a quantifier that is underinformative but logically true relative to its context ("some" after an "all"-type picture), and of one that is logically false relative to its context ("all" after a "some"-type picture); importantly, these effects should become evident at the quantifier itself rather than downstream, allowing us to see whether scalar implicature generation modulates ERPs at an immediate level.

Furthermore, the present study is conducted using Mandarin Chinese. Little research has been done on the processing of scalar implicatures in Chinese; all the existing psycholinguistic research I am aware of is on European languages (English: Nieuwland et al., 2010; Grodner et al., 2010; Huang & Snedeker, 2009; Tavano, 2010; Degen & Tanenhaus, 2010; Hunt et al., in prep.; French: Bott & Noveck, 2004; Noveck & Posada, 2003; Greek: Breheny, Katsos, & Williams, 2006; German: Degen, 2009; Dutch: De Neys & Schaeken, 2007; Italian: Foppolo, 2007), and even much of the theoretical literature in Chinese publications still operates mainly on English data (e.g., Fang, 2008; Zhu & Wang, 2007; but see Chi, 2000). There is a growing body of evidence that Chinese speakers process language information differently than speakers of Indo-European languages insofar as the integration of syntactic and semantic information in incremental sentence processing, for example, is concerned (Ye et al., 2006; Yu and Zhang, 2008;

Zhang et al., 2010); furthermore, the offline results of Wu and Tan's (2009) investigation of scalar implicature in Chinese indicate that Chinese speakers may have a stronger bias towards interpreting the scalar term *yǒu de* 有的 pragmatically than, for instance, English speakers do for "some".[4] Thus, it is worthwhile to broaden the coverage of experimental research on scalar implicature by seeing if Chinese speakers' computation of pragmatic implicatures differs from computation of comparable phenomena by speakers of other languages.

1.5 Related ERP findings: inference and quantification

Two additional ERP studies, although not specifically investigating scalar implicature, bear mention because of their conceptual similarities to the present study and their implications for the interpretation of my results. Pijnacker and colleagues (2011) investigated brain responses to statements in which a previously-assumed inference must be cancelled because of conflict introduced by a contextual exception. For example, the *modus ponens* shown in (8) implies that Lisa will wear contact lenses, whereas that in (9) does not:

8) a. Lisa has recently bought contact lenses.
   b. If Lisa is going to play hockey, then she will wear contact lenses.
   c. Lisa is going to play hockey.
9) a. Lisa seems to have lost a contact lens.
   b. If Lisa is going to play hockey, then she will wear contact lenses.
   c. Lisa is going to play hockey.

[4] In their study, adults rejected 89% of underinformative but logically correct "some" sentences. Rejection rates of 44% have been reported for English (Tavano, 2010); 59% and 63% for French (Bott & Noveck, 2004; Noveck & Posada, 2003); and 76% for Dutch (De Neys & Schaeken, 2007). It is difficult to make direct comparisons between these percentages, however, as ratings are susceptible to effects from methodological differences including instructions, filler types, alternative quantifiers or numbers available, and the set size being probed (see Degen & Tanenhaus, 2010a; Huang et al., 2010).

The critical condition involved a comparison between the sentence "Lisa will wear contact lenses" (examples translated from Dutch) in the context of (8), where it is congruent with the inference, and in the context of (9), where it is incongruent because an exception to the inference was introduced in (9a). In response to sentences in which the inference conflicts with the context, the investigators found a sustained negativity, centro-parietal in topography but with a different time course than the lexico-semantic N400 effect. The sustained negativity had a longer duration than the N400 effect, and unlike the N400 effect it lacked a clear peak. The investigators interpreted the sustained negativity as reflecting processing related to the revision of an already-computed inference in order "to incorporate an exception" (Pijnacker et al., 2011, p. 479). While the present study investigates different linguistic phenomena than that study, processing the underinformative sentence may nonetheless involve cancelling the already-computed scalar implicature ("*some* means *not all*") to arrive at the logical interpretation of the quantifier ("*some* means anything more than one"). It is not unreasonable to expect that this computation may resemble the one performed by participants in Pijnacker and colleagues' (2011) study.

Jiang and colleagues (2009) invested processing of the Mandarin universal quantifying adverb *dōu* 都. While not syntactically a quantifier (Li & Thompson, 1981), *dōu* is similar to the "all"-type quantifier *suǒyǒu de* 所有的 in the present study, in that it is only licensed in a context where its antecedent is a plural, preferably exhaustive, referent (such as "all the people"). In three experiments examining different sentential contexts and tasks, Jiang and colleagues (2009) compared brain responses to the quantifying adjective and following words in sentences such as the following (examples translated from Chinese):

10) a. After the consultation, the patients all gave up their bad habits.
    b. *After the consultation, the patient all gave up his bad habits.

When the participants' task was to perform an acceptability judgment (Experiments 1 & 2), the unlicensed adverb elicited a sustained positivity regardless of what sort of sentence structure it was embedded in, and the positivity continued to appear on subsequent words. The positivity did not appear when the task was merely to passively read the sentences for comprehension (Experiment 3). In addition, a sustained centro-parietal sustained negativity was elicited by the violation in all three experiments; during the acceptability judgment (Experiments 1 & 2) the negativity did not appear until the word following the quantifier and had a centro-parietal distribution, whereas during the passive reading experiment (Experiment 3) it appeared in the 600-900 ms time window following the quantifier itself and had a left-frontal distribution. The authors interpreted the sustained positivity as reflecting a problem in linking the quantifying adverb to a featurally inappropriate antecedent, and the sustained negativity as reflecting second-pass processing as the participants attempted to interpret the syntactically anomalous sentence.

1.6 Predictions for the present study

Because previous investigations using eye-tracking (Grodner et al., 2010; Tavano, 2010; Degen & Tanenhaus, 2010b) and self-paced reading (Breheny et al., 2006) have shown that scalar implicatures can elicit immediate effects when the quantifier is encountered, I predict that ERP responses for underinformative, pragmatically-violated stimuli ("some"-type sentences after "all"-type pictures) will differ from those for control stimuli. The precise nature of this effect, however, is not straightforward to predict. It is possible that an increased N400 effect[5] similar to

[5] N400 effects are well-attested in Chinese and generally have a comparable latency and topography to N400 effects in other languages. Ye and colleagues (2006) elicited a very early N400-like effect and argued that semantic processing may be sped up in Chinese, but later studies of semantic processing of Chinese (e.g. Ye et al., 2007; Yu & Zhang, 2008; Jiang & Zhou, 2009; Zhou et al., 2010), have elicited N400 effects with more typical latencies. The

that found on underinformative words later in the sentence (Nieuwland et al., 2010) will be obtained; such an effect may be directly due to the computation of scalar implicature and comparison of the interpretation derived from the implicature against the context given, or it may be related to the violated expectation of an "all"-type quantifier. It should be noted, though, that the critical words are repeated numerous times through the experiment, and repetition is known to reduce N400 amplitudes (Kutas & Federmeier, 2000), which could obscure effects. Another effect that may be elicited is a sustained negativity. Sustained negativities have been elicited by linguistic stimuli which cancel or cause conflict with a previously generated inference (Pijnacker et al., 2011; Baggio et al., 2008). If participants in the present study generate the *not all* reading of "some" at first and then attempt to override this meaning to integrate the sentence with the picture, this sort of sustained negativity may appear. On the other hand, given that Nieuwland and colleagues interpreted the effects they observed at the quantifier region as being consistent with an account that low-ability participants were suppressing implicatures, implicature suppression in this study might lead to a right-lateralized positivity in a time window near 350-400ms or a late frontal negativity. Their comparison, however, is not similar to the comparison that will be made in the present study (the authors compared "some" to another quantifier in contexts where no information was yet available about the correctness of the sentence, whereas we will be comparing "some" in correct contexts to "some" in a context where it is introducing a pragmatic violation), and the stimuli in Nieuwland et al. (2010) were not designed to test effects at this region, so it is not strongly predicted that such effects will be observed. I have no other *a priori* predictions about the effects of the pragmatic violation.

---

study by Ye and colleagues (2006) used auditory presentation, whereas the others cited all used rapid serial visual presentation (RSVP). The current study uses RSVP, and is currently being replicated in the auditory domain.

As for the logic violation, I predict that an N400-like effect may be elicited as in the pragmatic violations, since the quantifier is unexpected. It is possible, however, that the effect will be weaker or absent since the sentence is not necessarily incorrect when the quantifier is encountered (for instance, if a picture shows some chefs chopping cucumbers and some chefs chopping onions, a sentence beginning "In the picture, all…" may still be felicitously continued as "…all the chefs are wearing hats", assuming that they actually are wearing hats). It is also possible that the logic violation in this study could elicit a sustained positivity like that elicited by Jiang and colleagues (2009) in response to the syntactically unlicensed quantifying adverbial, given that the logic violation also involves a failed attempt to link an "all"-type term with a non-exhaustive antecedent—in that study it the inappropriate antecedent was a non-plural noun, whereas in the present study it is either set of actors in a picture, under a context in which these actors do not exhaust the set of actors shown in the picture. On the other hand, because the linguistic phenomena examined in the two studies are different (Jiang and colleagues' (2009) manipulation is a syntactic one rather than a semantic one, whereas ours does not involve any syntactic ill-formedness), it is just as possible that different effects be elicited.

As described above, Nieuwland and colleagues (2010) found opposite N400 effects for participants with low and high pragmatic abilities, such that when all participants were analyzed together there was no effect visible in the N400 time window. The present study includes no test of participants' pragmatic abilities and thus does not allow for the grouping of participants in this way.[6] The reversal of the N400 effect for low-ability participants, however, was shown to be due to a lexico-semantic N400 in the opposite direction based on the extent to which the critical

[6] Using the dependent variable as a grouping variable—i.e., grouping participants based on the ERP they demonstrate—is statistically questionable (see, e.g., Kliegl et al., 2011).

words were primed by preceding words. In the present investigation, the critical words are the quantifiers themselves, which are not content words and which should be relatively unaffected by lexico-semantic relatedness to other words in the sentence, even for low-ability participants (simply because these words have little lexico-semantic, or referential, content, and thus the amount to which they are semantically related to other items in the trial is necessarily constant across conditions and items). Noveck and Posada's (2003) participants were also split into two groups based on their behavioral performance, but these groups showed no significant differences in ERP responses at either the critical word or the quantifier (see Noveck & Posada, 2003, note 4). Thus, it is not necessarily expected that our participants will show a similar bimodal distribution of effects.

2 Methods

2.1 Participants

In this thesis I report a preliminary analysis of data that was collected from nine right-handed Mandarin native speakers (6 female, age range 18-24, mean 19.8) from mainland China who were students at the University of Kansas. I also have data from ten additional participants whose recordings show substantial ocular and muscle artifacts compared to the nine reported here; because the data collected from these participants requires additional artifact correction and processing beyond the scope of the preliminary analysis in this thesis, I restrict the present report to the nine participants describe above.[7]

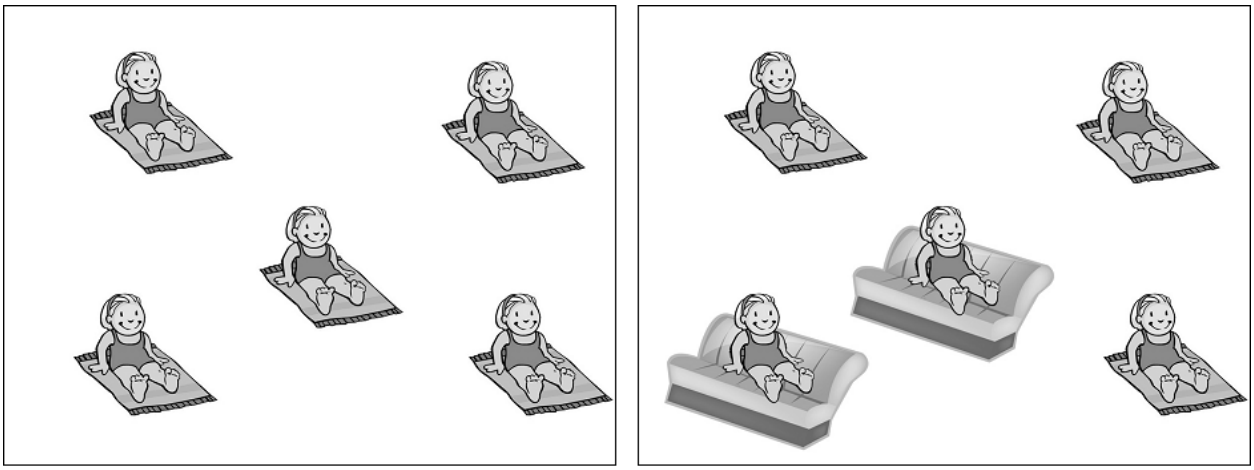
Many of the participants were bilingual in Standard Mandarin and a local language or dialect, but all participants reported that Standard Mandarin was the language they acquired earliest and the one they use most. All had normal or corrected-to-normal vision and were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). Participants were recruited through fliers, university e-mail, and word of mouth. They received payment, and all methods for the study were approved by the Human Subjects Committee of Lawrence (HSCL) at the University of Kansas.

[7] It is likely that the unusually high occurrence of artifacts was due to the experimental design. The pictures displayed for 4 seconds on each trial were quite bright, and participants performed the task in a darkened room; the bright pictures made some participants uncomfortable, resulting in frequent eye artifact; in addition, the inter-trial interval was made short (1000 to 1800 ms) to keep the recording session from being excessively long, but this resulted in more eye blinks during the sentence presentation. Furthermore, since only 20% of sentences were followed by probes (either comprehension questions or acceptability judgments), some participants reported feeling bored and sleepy during the experiment, which resulted in alpha artifact. I am currently replicating this study, however, using adjusted procedures and a more comfortable auditory presentation, to gather more robust results with regards to the research question.

2.2 Materials

One hundred sixty picture sets were created for the critical trials. Each set included three to five actors or items. In the "all"-type picture from each set, all of the actors were interacting with identical objects (for instance, four girls all sitting on blankets, or five baskets all holding pumpkins). In the "some"-type picture from each set, a subset of the actors (at least two) were interacting with that object, and the rest interacting with a different object (for instance, some girls sitting on blankets and some on sofas, or some baskets holding pumpkins and some bananas). The placement of the actors within the image, and the relative locations of actors with different items in the "some"-type pictures, was allowed to vary randomly across sets. All the pictures were black-and-white cartoons or line drawings sized 1024x768 pixels and with minimally complex backgrounds. Care was taken to limit pictures to those portraying plausible events. Sample pictures of each type are shown in Figure 1. The base materials for the pictures were taken from freely available clipart from two published databases (Bonin et al., 2003; and the International Picture Naming Project: Szekely et al, 2004), and Google Images, and further edited using Adobe Photoshop, the GNU Image Manipulation Program, and Microsoft Paint by two paid graphic arts students from Peking University and by the author.

For each picture set, "all" and "some" sentences were written to match the "all" and "some" pictures (see Figure 1). Each sentence adhered to the structure exemplified in Table 1, with the exception that some sentences had no tail (region 7) in cases where no natural continuation was possible. All critical sentences used either the quantifier *yǒu de* (有的), "some", or *suǒyǒu de* (所有的), "all". The Mandarin quantifier *yǒu de* is an inherently scalar term (see, e.g., Chi, 2000) and is expected to elicit scalar implicatures. Wu and Tan's (2009) adult participants tested in a similar paradigm as the current study's reported a pragmatic interpretation



A **B**

i)	图片里，/	所有的/	女孩/	都/	正坐在/	毯子上/	晒太阳。
	In this picture,	all of the	girls	DOU	sitting on	blanket POS.	suntanning
ii)	图片里，/	有的/	女孩/	/	正坐在/	毯子上/	晒太阳。
	In this picture,	some of the	girls	-	sitting on	blanket POS.	suntanning

"In this picture, some/all of the girls are sitting on blankets suntanning.

Figure 3: a) "all"-type picture, with corresponding "all"-type sentence (i); b) "some"-type picture, with corresponding "some"-type sentence (ii)

Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7
"in this picture",	Quantifier	Subject	（都）	Verb + Aspect	Object	Tail

Table 1: Sentence structure used in the experiment

in 89% of trials. The stimuli in the present study occur in a context that is neutral for scalar implicature generation. It has been demonstrated (for reviews see Katsos & Cummins, 2010; Noveck & Sperber, 2007) that context plays a significant role in both whether the final interpretation of an utterance includes a scalar implicature, and in the computational processes undergone to generate (or not generate) it. The crucial difference is between upper-bounding contexts, in which what is relevant to the discourse is the upper part of the scale (e.g., whether *all* the characters are engaging in the activity), and lower-bounding contexts, in which what is relevant to the discourse is the lower part of the scale (e.g., whether *any* of the characters are engaging in the activity) (Katsos & Cummins, 2010, p. 285). Because the context of the sentences in the experiment is merely an expectation that they be consistent with pictures, both upper- and lower- bounded descriptions can be accurate; for example, an accurate description of a picture in which five girls are all sitting on blankets could be an existential (presentational) sentence "There are girls sitting on blankets" (有女孩子正坐在毯子上), satisfying the lower bound, or a typical subject-verb-object sentence "All of the girls are sitting on blankets" (所有的女孩都正坐在毯子上), satisfying the upper bound; the context itself does not specifically require a lower- or upper-bounded description. A description of the lower bound, however, is only felicitous with an existential/presentational quantifier, not with a partitive.[8] The Mandarin quantifier *yǒu de* in this sentence structure, however, is considered a modifier in the form of a relative clause (Xie, 2003) or a DP-modifying partitive (Tsai, 2003, 2004), and thus the stimulus sentences used in the present study, which all use *yǒu de*, can be considered subject-object-verb

[8] There is also evidence in English that the partitive "some of" (auditory "summa") has a stronger pragmatic interpretation than bare "some", which is presumably more existential/presentational (Grodner et al., 2010; Degen & Tanenhaus, 2010a).

sentences whose subjects are modified by restrictive relative clauses or partitive quantifiers; either way, they are not existential sentences. In other words, they are more like the second example, which is expected to accurately describe the upper bound in order to be felicitous. Since the truth or falsehood of "all" is conversationally relevant in an upper-bounding context, this is a context in which a quantifier "some" is expected to generate a scalar implicature and, in turn, to be underinformative.

Verbs in the critical sentences were in either the progressive aspect (indicated by the post-verbal aspect marker *zhe* 着 or the preverbal aspect markers *zài* 在, *zhèng* 正, or *zhèngzài* 正在) or the perfective aspect (indicated by the aspect marker *le* 了), or were preceded by a coverb indicating prospective aspect (either *yào* 要, "is about to", or *zhǔnbèi* 准备, "is preparing to"). "All"-type sentences included the mandatory adverbial *dōu* 都 before the verb (see, e.g., Li & Thompson, 1981; Jiang et al., 2009). Subjects and verbs were allowed to be repeated across sets. The sentences were written with the help of a paid linguistics student from Peking University who was a native speaker of Mandarin.

By crossing "some"- and "all"-type pictures with "some"- and "all"-type sentences, a fully factorial 2x2 design was created (see Table 2). "Some"-type sentences preceded by "some"-type pictures were correct and felicitious descriptions of the pictures, whereas those preceded by "all"-type pictures were logically correct but infelicitious, or underinformative, descriptions because they violated a scalar implicature. "All"-type sentences preceded by "all"-type pictures were correct and felicitous descriptions of the pictures, whereas those preceded by "some"-type pictures were logically incorrect descriptions. In this way, pragmatically infelicitous and logically incorrect sentences could each be compared against correct controls that differed only

in preceding picture context but were completely matched in lexical content and syntactic structure.

	"all" sentence	"some" sentence
"all" picture	**Correct "all" x 40**	**Scalar violation x 40**
"some" picture	**Logic violation x 40**	**Correct "some" x 40**

Table 2: Conditions included in the 2x2 design

Additionally, 148 pictures were created for use as fillers. The filler pictures met the same specifications as the critical trials, except that some of them depicted intransitive events (e.g., cats sleeping or birds singing) and thus the sentences did not include objects. Thirty-seven of these fillers were "some"-type pictures and were paired with matching, felicitous "some"-type sentences, and thirty-seven were "all"-type pictures and were paired with matching, logical "all"-type sentences. Another thirty-seven of the pictures (19 "all"-type and 18 "some"-type) were paired with sentences that had an appropriate quantifier but an object that did not match any of the objects shown in the picture. The final thirty-seven (18 "all"-type and 19 "some"-type) had an appropriate quantifier but a verb that did not match the activity shown in the picture. Several of these had verbs that yielded semantically anomalous sentences (e.g., “all the scientists are *planting squirrels”), whereas most had verbs that were semantically plausible but not congruous with the picture. The filler sentences all included quantifiers that were not used in the critical sentences (*yǒu xiē* 有些, "some"; *yǒu yì xiē* 有一些, "a few"; *shǎo shù* 少数, "a small number of"; *xǔ duō* 许多, "many", *dà bùfen de* 大部分的, "most of"; *quánbù* 全部, "all"; *quánbù de* 全部的, "all of"), or classifier phrases in place of quantifiers (*jǐ* 几 + classifier, "a few"; *hǎo jǐ* 好几 + classifier, "many"; or *měi* 每 + classifier, "every"). The set of mismatching picture-sentence

fillers were included to distract participants from the quantifier manipulation in the critical sentences, and the remaining matching fillers were included to keep the proportion of acceptable and unacceptable sentences to at least 50% during the experiment. (For participants who accept the infelicitous, underinformative stimuli, the number of acceptable trials exceeds the number of unacceptable ones.)

2.3 Procedure

Participants were seated comfortably in a dimly-lit room. They were about 1 meter in front of a 41-cm CRT monitor. The pictures and sentences were presented at the center of the screen using the Presentation software package (Neurobehavioral Systems, http://www.neurobs.com). Each trial began with a fixation point presented for 500 ms, followed by a picture which remained on the screen for 4000 ms. The picture was followed by a fixation point of variable duration (for a random amount of time between 500 and 1500 ms), after which the sentence was presented region by region using the rapid serial visual presentation paradigm. Regions were presented by way of a variable presentation procedure (see, e.g., Nieuwland et al., 2010), whereby each region was presented at a base rate of 425 ms per region, plus 80 ms for each character more than 3 in the region; because the critical quantifiers were all three characters or less, their presentation durations do not differ across or within conditions. The interstimulus interval was 400 ms for all regions.[9] Twenty percent of trials were followed by comprehension questions or acceptability judgments (see below), which were presented in the screen for 5000ms

[9] An 800-millisecond stimulus onset asynchrony (400 ms word presentation, 400 ms interstimulus interval) has been found to be natural and comfortable for Chinese readers in previous studies (Li & Zhou, 2010; Jiang & Zhou, 2009; Ye & Zhou, 2008), but the regions used in the present study were, on average, longer than the regions used in those studies, and pilot participants reported the variable presentation rate described above to be the most comfortable.

or until the participant's response. Each trial was followed by a blank screen for 1500 ms before the start of the next trial. The experiment was divided into six blocks of about 50 sentences each, and participants were allowed to take breaks between the blocks. Participants were instructed not to blink during the presentation of the sentences.

Participants were asked to perform a mixture of acceptability judgments and comprehension probes. In ten percent of trials, after the sentence was presented, a question that probed information about the picture and was irrelevant to the sentence was also presented (e.g., after the sentence "In this picture, some of the girls are sitting on blankets", the comprehension question "Are the girls wearing swimsuits?" was presented). In an additional ten percent of trials, the sentence was followed instead by an acceptability judgment (the question *du ìbú du ɐ?* 对不对, meaning "Is that correct?"). Participants were not given instructions about how to judge sentences before the experiment unless they asked; if they asked, they were instructed to judge, based on their own intuition, whether the sentence was consistent with the picture and described it appropriately. The comprehension questions were included to prevent participants from being able to adopt a strategy of only paying attention to the quantifiers and the number of objects in a picture, and the acceptability judgments were included to force participants to pay attention to the sentence rather than just trying to remember the picture. Within the critical trials, 12 of the 16 comprehension questions were on correct, felicitous sentences (6 each for correct "all" and correct "some") conditions and only 4 out of 16 acceptability judgment prompts were; within the fillers, this pattern was reversed. Participants responded to the comprehension questions and acceptability judgment prompts using the left and right mouse buttons.

The experimental sentences were divided into four lists according to a Latin square design, such that every sentence appeared once in each condition across lists but no sentence or

picture was repeated within a list. The lists and fillers were randomized for each participant. The first block of the experiment was preceded by a practice block of seven trials which followed the same presentation procedure as the main experiment but did not include any quantifier-related violations. The practice sentences included some sentences with existential quantifiers (e.g., "in the picture there are…", 图片里有。。。) and some without quantifiers (e.g. "the dogs in the picture are…", 图片里的小狗。。。). Feedback was given for behavioral responses in the practice block, but not in the main experiment. The entire recording took about 75 minutes per participant, not including setup time.

2.4 EEG recording and data acquisition

The EEG was continuously recorded using an elastic electrode cap (Electro-Cap International, Inc.) containing 32 Ag/AgCl scalp electrodes organized in a modified 10-20 layout (midline: FPz, Fz, FCz, Cz, CPz, Pz, Oz; lateral: FP1/2, F7/8, F3/4, FT7/8, FC3/4, T3/4, C3/4, TP7/8, CP3/4, T5/6, P3/4, O1/2). Polygraphic channels were placed at the left and right outer canthi for monitoring horizontal eye movements, above and below each eye for monitoring blinks, and on the left and right mastoids. The left mastoid served as a reference during data acquisition and AFz served as the ground. Impedances for scalp electrodes and mastoids were kept below 5 kΩ. The recordings were amplified by a Neuroscan Synamps2 amplifier (Compumedics Neuroscan, Inc.) with a bandpass of 0.01 to 200 Hz and digitized at a sampling rate of 1000 Hz.

2.5 Data analysis

EEG waveforms were inspected manually to reject artifacts. Trials containing excessive eye or muscle artifact or alpha activity were excluded from the analysis; participants with less than 64% good trials were also excluded from the analysis. Participants were not excluded from the analysis on the basis of their behavioral responses, as these would only account for 20% of trials and were regardless deemed not to be a good reflection of how attentive they were to the critical violations.[10]

For the remaining participants, the EEG was re-referenced to the average of both mastoids, and ERPs for each condition were computed by averaging artifact-free trials over epochs of 200 ms before to 1000 ms after the presentation of the quantifier. ERPs were baseline-corrected using a 200-ms prestimulus baseline and subjected to a 30 Hz low-pass filter. Filtered data was used for the statistics reported in this paper. Mean voltage amplitudes were calculated over the time windows of interest (see section 3.2, "ERP results") for each electrode. Electrodes were assigned to regions on the basis of anteriority (two levels) and laterality (three levels), yielding the following nine regions: left anterior (FP1, F7, F3, FT7, FC3), midline anterior (FPz, Fz, FCz), right anterior (FP2, F8, F4, FC4, FT8), left posterior (T3, TP7, T5, C3, CP3, P3, O1), midline posterior (Cz, CPz, Pz, Oz), and right posterior (T4, TP8, T6, C4, CP4, P4, O2). Region averages were calculated using this layout.

[10] As for the acceptability judgments, for the scalar violations there is no "right" answer, so it would be impossible to exclude participants on the basis of behavioral performance in this condition; furthermore, across conditions, correct trials were occasionally judged as incorrect by participants who found the wording of the sentence awkward regardless of the status of the quantifier. As for comprehension questions, many of these probed relatively non-salient details about the picture (e.g., "do the moose have tails?"), and a participant's failure to remember these irrelevant details accurately on a given trial does not mean that the participant did not process the quantifier actively.

The mean voltage amplitudes in each time window were subjected to a 2 (sentence type: pragmatic, logic) x 2 (violation: violation, no violation) x 2 (anteriority: anterior, posterior) x 3 (laterality: left, midline, right) repeated-measures ANOVA. The Greenhouse-Geisser correction was applied for violations of sphericity. Only significant effects are reported below.

3 Results

3.1 Behavioral results

Data from one participant was lost due to a data logging error, leaving eight participants for the behavioral analysis. Here I report the participants' judgments of the pragmatically underinformative sentences (those with the quantifier "some", following "all"-type pictures). Each participant performed an acceptability judgment on 6 of these sentences (out of 40), yielding a total of 48 judgments. Across participants, 32 of these sentences (67%) were judged as correct, indicating a logical judgment; in comparison, 90% of logically incorrect sentences were rejected. The difference was significant by participants, $t(7) = 4.252$, $p = .004$, indicating that participants accepted pragmatically infelicitous sentences more often than logically incorrect sentences.[11]

Participants were further subdivided based on their response patterns: those who made 5 or more logical responses (1 or fewer pragmatic responses) to the underinformative trials were classified as logical responders, those who made 5 or more pragmatic responses (1 or fewer logical responses) were classified as pragmatic responders, and those who made 2 to 4 logical responses (no more than 4 responses of a given type) were classified as inconsistent responders. Five participants met the criteria to be considered logical responders, while only one participants was a pragmatic responder and two were inconsistent responders.[12] These results suggest that

[11] When behavioral data from participants whose EEG data was not used in the present report, the same pattern holds (75% acceptance of pragmatically infelicitous sentences, and 88% rejection of logically incorrect sentences) and the difference is still significant, $t(18) = 7.065$, $p < .001$.

[12] Using slightly more lax criteria (4 or more logical responses for logical responders, 4 or more pragmatic responses [2 or fewer logical responses] for pragmatic responders, and 3 logical [3 pragmatic] responses for inconsistent responders), the pattern of data was roughly the same: 5 logical responders, 2 pragmatic, and 1 inconsistent. Furthermore, the same pattern holds in the analysis including data from all subjects, both with the

participants generally made consistent judgments about underinformative sentences, a finding in agreement with results from similar tasks (Tavano, 2010; Noveck & Posada, 2003).

3.2 ERP results

Grand average waveforms for all conditions, across nine participants, are shown in Figure 2. A clear N1 – P2 pattern, common in paradigms using visual presentation, is apparent in all conditions. Visual inspection of the waveforms reveals that the word "all" (both blue lines) elicited a larger negativity than the word "some" in the 300-500 ms time window over frontal electrodes, regardless of the condition in which it appeared. More importantly, a broadly distributed sustained negativity appears in both pragmatic (underinformative) and logical (patently false) violations—solid lines, relative to dashed lines—in the time window of about 300 to 900 ms after the onset of the quantifier.

To select appropriate time windows for statistical analysis and to quantify how sustained the effect was, I performed factorial repeated-measures ANOVAs (see section 2.5, "Data analysis") over consecutive 50-ms time windows from 150 ms to 900 ms. The results are summarized in Table 3. Effects of interest are those that include Violation and/or Type as factors, as these are the only effects likely to have been caused by the experimental manipulation. Within the effects of interest, two clearly different patterns of effects emerged: from 150 to 500 ms there was a consistent interaction of Anteriority and Violation, whereas from 600 to 900 ms there was a consistent interaction of Laterality, Type, and Violation. Based on these findings I selected

---

strict criteria (yielding 13 logical responders, 1 pragmatic responder, and 4 inconsistent responders) and the lax criteria (14 logical, 4 pragmatic, 1 inconsistent).

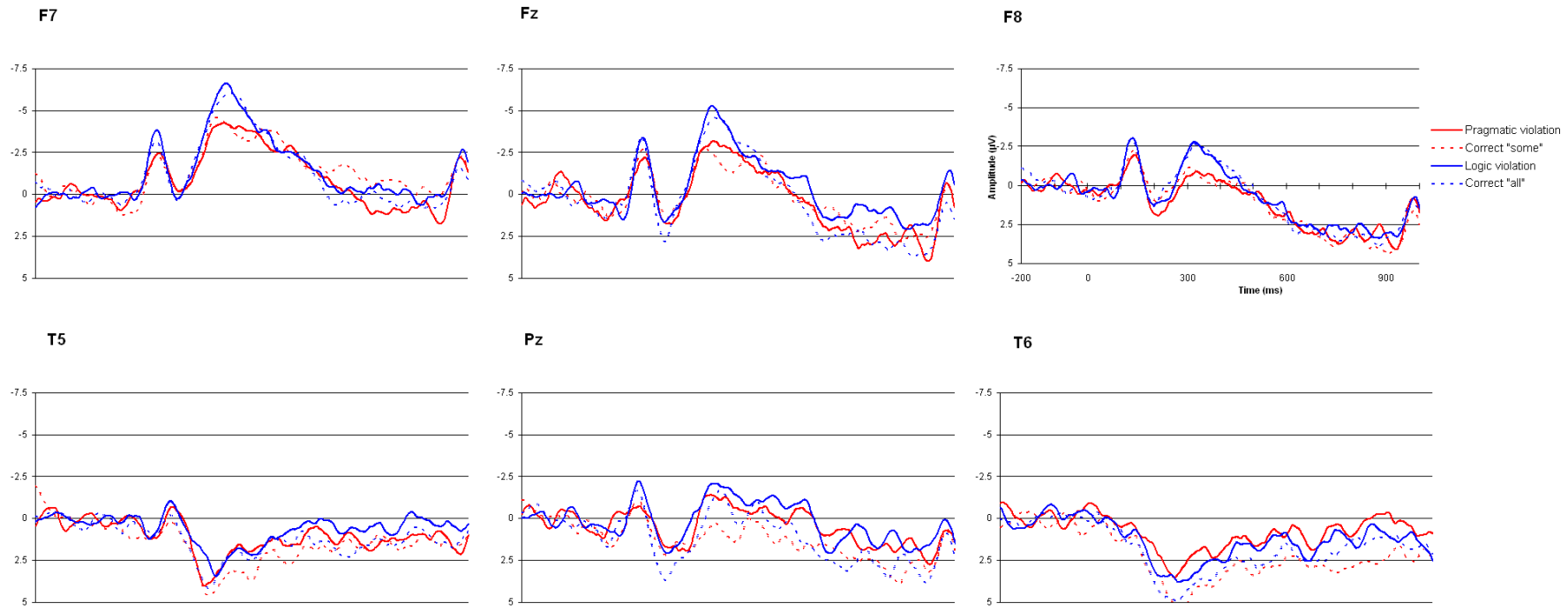


Figure 4: Grand average waveforms at six selected electrodes (one per region). A 15 Hz low-pass filter was applied to the data for plotting purposes for these waveforms. Negative is plotted upwards.

	150-200ms	200-250ms	250-300ms	300-350ms	350-400ms	400-450ms	450-500ms	500-550ms	550-600ms	600-650ms	650-700s	700-750s	750-800ms	800-850ms	850-900ms
ant		(0.051)	0.003	0.002	0.011	0.043									
lat	0.007	0.024	0.001	0.002	0.012	0.009	0.032	0.001	0.006	0.006	0.003	0.001	0.008	0.007	0.002
type				0.005	0.04										
vio					**(0.068)**										
ant*lat	(0.099)		0.001	0	0	0.001	0.003	0.009	0.003	0.005	0.011	0.018	0.027	0.035	0.027
ant*type				0.035											
lat*type															
ant*lat*type							0.047	0.045		0.014					
ant*vio	**0.032**	**(0.058)**	**0.02**	**0.032**	**0.001**	**(0.06)**	**(0.068)**				**(0.053)**				
lat*vio												**0.005**		**(0.078)**	
ant*lat*vio									**(0.094)**						**(0.06)**
type*vio															
ant*type*vio													**(0.096)**		
lat*type*vio										**0.018**	**(0.096)**	**0.01**	**0.01**	**0.024**	**0.046**
ant*lat*type*vio															

Table 3: Greenhouse-Geisser-corrected p-values for the ANOVAs conducted over consecutive 50-ms time windows. Marginal values ($0.1 < p \leq 0.05$) are shown in parentheses; values above 0.1 are not shown. Rows that are bolded with gray background indicate effects that are of interest because they include the Violation factor, meaning that the effect was modulated by the experimental manipulation. Rows with red text indicate effects that include a Type*Violation interaction, meaning that there were different effects of violation for pragmatic than for logical conditions.

three time windows for statistical analysis: 150-300 ms, 300-500 ms,[13] and 600-900 ms. In the material that follows, I report all significant *F*-statistics, but only attempt to interpret that are of interest and do not discuss effects that only involve topographical factors.

3.2.1 150-300 ms

The repeated-measures ANOVA in this time window revealed significant main effects of Anteriority ($F(1,8) = 6.346$, $p = .036$) and Laterality ($F(1.786,14.287) = 7.242$, $p = .008$), as well as an Anteriority*Laterality interaction ($F(1.342,10.732) = 4.903$, $p = .041$). More importantly for the present study, there was also a significant Anteriority*Violation interaction ($F(1,8) = 11.627$, $p = .009$). While violating sentences, regardless of violation type, were marginally more negative than controls over posterior sites (1.122 μV versus 1.929 μV, $t(8) = -1.889$, $p = .096$), they did not differ significantly from controls at anterior sites (-0.244 μV versus -0.3413, $t(8) = 0.217$, $p = .833$). The scalp distribution of the effect is illustrated in Figure 3. No other effects approached significance.

3.2.2 300-500 ms

The ANOVA in this time window revealed significant main effects of Anteriority ($F(1,8) = 9.264$, $p = .016$), Laterality ($F(1.626,13.005) = 8.171$, $p = .007$), and Type ($F(1,8) = 7.526$, $p = .025$), as well as a significant of Anteriority*Laterality

[13] Although roughly the same effects were present for the 150-300 ms and 300-500 ms time windows in the consecutive time-window ANOVAs, I divided them into two time windows based on, first of all, qualitative differences in the waveforms (the 150-300 ms time window included the N1 and P2 components), and secondly, the fact that 300-500 ms coincides with the traditional N400 time window (Lau, 2008; Hagoort et al., 2003; Kutas & Federmeier, 2000).

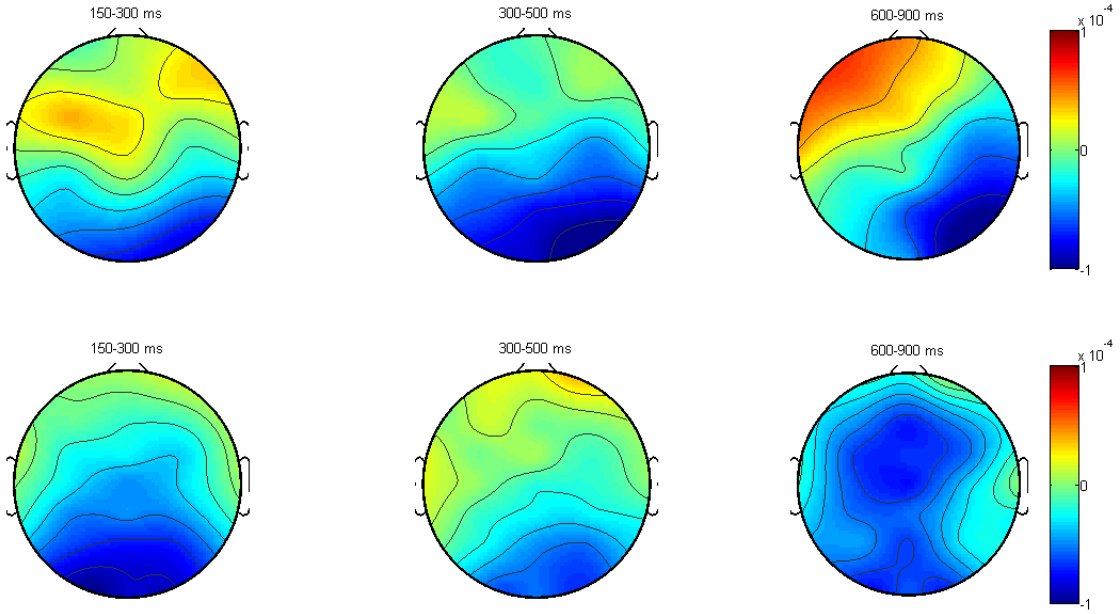


Figure 5: Scalp distributions of the pragmatic effect (underinformative – correct "some", top row) and logic effect (logically false– correct "all", bottom row) in the three time windows of interest.

(F(1.864,14.912) = 19.589, p > 0.001) and a marginal interaction of Anteriority*Laterality*Type (F(1.855,14.836) = 1.514, p = .078).

The main effect of Type indicates that "all"-type sentences elicited greater negativities than "some"-type sentences regardless of whether or not they were correct (-1.403 μV versus -0.568 μV, t(8) = 2.743, p = .025), as can be seen clearly in the waveforms, and the marginal Anteriority*Laterality*Type interaction suggests that the effect was broadly distributed over anterior electrodes but confined to the left and central regions over posterior electrodes. Collapsing across the Violation factor, the effects of Type at anterior electrodes are as follows: Left: t(8) = 2.174, p = .026; Midline: t(8) = 2.297, p = .051; Right: t(8) = 2.358, p = .046;

whereas over posterior electrodes the effect was marginal over left electrodes and increasingly weaker over midline and right electrodes: Left: $t(8) = 1.928$, $p = .090$; Midline: $t(8) = 1.677$, $p = .132$; Right: $t(8) = 0.755$, $p = .472$. Of greater relevance to the present study, there was also a significant Anteriority*Violation interaction ($F(1,8) = 12.196$, $p = .008$), due to the fact that violating sentences of each type elicited a significantly greater negativity than controls, regardless of type, over posterior regions (-0.298 μV versus 0.552 μV, $t(8) = -2.866$, $p = .021$) but not over anterior regions (-2.081 μV versus -2.113 μV, $t(8) = 0.072$, $p = .944$). No other effects approached significance.

The fact that the Type factor did not interact with the Violation factor at all in this time window suggests that both violation types elicited statistically identical N400 effects. An exploratory region-by-region analysis using paired-samples t-tests, however, suggests that the posterior N400 effect only reached significance in the pragmatic violation (all ps < .031), whereas it did not reach significance in any regions in the logic violation (all ps > .192). The different findings for these two analyses may reflect that there was insufficient data to detect this interaction in the repeated-measures ANOVA.

3.2.3 600-900 ms

The ANOVA in this time window revealed a significant main effect of Laterality ($F(1.765,14.124) = 10.213$, $p = .002$) and a significant Anteriority*Laterality interaction ($F(1.469,11.752) = 6.487$, $p = .018$). More importantly, there was a marginal Laterality*Violation interaction ($F(1.487,11.893) = 3.175$, $p = .089$) and a significant Laterality*Type*Violation interaction ($F(1.85,14.802) = 11.399$, $p = .001$).

Post-hoc paired-samples t-tests revealed that the pragmatic (underinformativeness) effect only reached significance in the right posterior region ($t(8) = -4.112$, $p = .003$), whereas the logic effect did not reach significance in any regions and were barely marginal over the midline ($t(8) = -1.975$, $p = .084$). The lateralization of the pragmatic effect was further confirmed by examining the size of the difference waves: the underinformative violation elicited the largest negativity, relative to the correct condition, in the right region, whereas substantial negativities were not elicited on left and midline electrodes (left: 0.502 μV; midline: -0.019 μV; right: -0.832 μV). The effect on right electrodes was significantly more negative than that on midline and left electrodes ($t(8) = -3.151$, $p = .014$; $t(8) = -3.962$, $p = .004$, respectively), while the effect did not significantly differ between left and midline electrodes ($t(8) = 1.840$, $p = .103$). The logically false violation, on the other hand, elicited the largest negativity over the midline (left: -0.851 μV; midline: -1.225 μV; right: -0.615 μV), although none of these negativities individually was significant. Thus, the Laterality*Type*Violation interaction indicated that the pragmatic violation elicited a right-lateralized negativity, while the logic violation did not elicit a significant negativity (and the non-significant negativity elicited showed a more broadly distributed topography than that of the pragmatic violation). These effects are plotted in the right-hand column of Figure 3.

4 Discussion

The present study measured brain potentials evoked by Chinese sentences that are either pragmatically or logically inconsistent with a preceding context. It used a picture-sentence verification design (Wu & Tan, 2009; Tavano, 2010), addressing conceptual and methodological limitations that were present in the only two previous ERP studies of scalar implicature generation (Nieuwland et al., 2010; Noveck & Posada, 2003). In the behavioral task, I found that participants tended to be consistent in their judgments of underinformative sentences, and that most of them judged such sentences as correct. In the ERP analysis, I found that pragmatic violations elicited broad posterior negativities in early time windows (150-300 ms and 300-500 ms) and a right broad posterior negativity in a late time window (600-900 ms). I discuss the behavioral and electrophysiological findings below.

4.1 Behavioral findings

The proportion of logical responses to underinformative sentences in the present study (67%) was greater than that observed in most similar online studies (56% in Tavano, 2010; 52% in Degen & Tanenhaus, 2010b; 40.7% in Bott & Noveck, 2004; approximately 37% in Noveck & Posada, 2003; approximately 24% in De Neys & Schaeken, 2007) and far greater than that observed in the only other Chinese study to test similar materials (11%, Wu & Tan, 2009). The only other study I know of that has elicited a logical response rate this high is the one conducted by Foppolo (2007); while her study examined scalar implicatures elicited by the coordinator "or" rather than the quantifier "some", it nevertheless included an implicature-violating condition, which was judged logically (i.e., judged as "true") 77% of the time in non-downward-entailing

contexts, which are assumed to be conducive to scalar implicature generation, and 90% of the time in downward-entailing contexts, which are assumed to hinder scalar implicature generation.

The high likelihood of logical responses in the present study may be due to the nature of the filler sentences used: this study included sentences with highly salient semantic violations (describing objects or activities that were not present in the pictures), so participants may have decided that underinformative sentences are relatively acceptable in comparison with those sentences.[14] In online acceptability ratings collected using a 7-point Likert scale during a follow-up to this study (see Appendix I), underinformative sentences were rated significantly better than sentences with objects that did not match the picture, and significantly worse than completely correct sentences; a similar distribution of ratings has been found when comparing correct and lexico-semantically violated sentences to sentences with a syntactic quantification error (specifically, misuse of the Mandarin "all"-like adverbial *dōu* 都; Jiang et al., 2009, Experiment 1). Consistent with this interpretation, some participants reported after the experiment that they judged underinformative sentences as incorrect early in the session but began considering them correct later, as they encountered worse sentences from the fillers; unfortunately, because of the small number of judgments collected and the ability of a single judgment (even one based on information other than the quantifier, such as a participant feeling that the "tail" of a sentence

[14] In comparison, Noveck & Posada (2003) included no fillers in their study (although the sentences in the patently false condition involved semantically incongruous sentence-final words), perhaps drawing more attention to the quantification aspect of the sentence. De Neys & Schaeken's and Bott & Noveck's fillers included a small number of semantically incongruous words, but most of the violations were violations of quantification (e.g., "all trees are elms"). While Tavano's (2010) study included fillers, none of the fillers given in her examples include semantic anomalies. Foppolo (2007) does not include a description of the fillers used in that study. Nieuwland and colleagues (2010) used only true and semantically non-anomalous fillers, and because their participants read sentences passively, no judgment data is available from that study.

was awkward), no clear pattern emerged when I attempted to compare data from the first and second halves of the experiment.

Another factor contributing to the preponderance of logical responses in the present study may be the cognitive demands placed on participants by the task and the design: since participants sometimes had to answer irrelevant comprehension questions about the pictures, they were forced to maintain a representation of the picture in working memory, which may have made the task more cognitively demanding than tasks requiring only acceptability judgments. De Neys and Schaeken (2007) have shown that increased cognitive load increases the rate of logical responses given to underinformative sentences in an acceptability judgment task. The only other study to elicit a logical response rate as high as this one, Foppolo (2007), involved a task that also should have involved substantial cognitive resources (making a complicated logic judgment about four pictures simultaneously while trying to remember a previously-presented "if-then" statement consisting mostly of nonsense words—the task was so difficult that reaction times ranged from eight to twelve seconds); while that study did not specifically manipulate the amount of cognitive load, its results and the present results are nevertheless consistent with the conclusion of De Neys and Schaeken (2007), especially when compared with the acceptability judgment results of other studies, most of which used less cognitively demanding tasks.

In the follow-up to this study (Appendix I), the cognitive load required for the task was substantially lessened (all sentences were followed by a rating scale, and none were followed by comprehension questions probing irrelevant material), but underinformative sentences were still rated significantly higher than any other type of violation, suggesting that the high acceptance rate in the present study may have been more due to the influence of the fillers than to the cognitive load imposed by the task. Such an interpretation is convergent with the findings of

Degen and Tanenhaus (2010a), who found that the types of other quantifiers or numbers present in the stimulus set influenced participants' interpretations of "some" and "some of"; the results of the present study extend the results of that study by identifying another feature of the experimental context that may modulate judgments of underinformative sentences.

The other behavioral finding of note in the present study was that participants tended to be internally consistent in their evaluation of pragmatically underinformative sentences. This replicates the behavioral results of previous studies using underinformative stimuli (Noveck & Posada, 2003; Tavano, 2010). Tavano and Kaiser (2010), however, point out that both groups of participants, at least in their experiment, are nonetheless aware of the scalar implicature, and that the participant grouping may be an artifact of the experimental context and participants' conscious decisions to respond the way they thought the experimenters wanted them to respond; furthermore, in some cases there are no differences in online processing measures between groups (ERPs in Noveck & Posada, 2003) or trivial differences (eye movement measures in Tavano, 2010). For these reasons, I do not believe the finding of response consistency in participants is particularly meaningful for the current study, and has little bearing on the overall predictions or interpretations of the study.

4.2 ERP findings

4.2.1 Immediacy of implicature processing

The negativity elicited by violating stimuli in the 150-300 and 300-500 ms time windows did not replicate the late frontal negativity observed in Nieuwland and colleagues' (2010) exploratory analysis on quantifiers. Rather, it seems to be a classic N400 effect, based on its latency (N400 effects typically appear around 300-500 ms, and may appear earlier on critical

words that are highly repeated; see Renoult & Debruille, 2011), its centro-parietal distribution, and its elicitation by stimuli that are unexpected and are incongruous with the context (Lau, 2008; Kutas & Federmeier, 2000). The finding that pragmatically underinformative stimuli elicit an N400 effect replicates and extends the results of Nieuwland and colleagues' (2010) analysis on object words that introduce underinformativeness; the present design demonstrates that this underinformativeness effect can emerge the moment the quantifier is encountered and that it is independent of lexico-semantic relations and real-world knowledge. The latency of this effect is comparable to that of N400 responses to lexico-semantic anomaly or incongruity (for reviews see Lau, 2008; Kutas & Federmeier, 2000), which suggests that quantification is processed the moment a quantifier is encountered and that this processing occurs at an immediate level (or, at least, as quickly as the processing of lexico-semantic content words). The present study is, to the best of my knowledge, the first ERP study to demonstrate immediate pragmatic effects when a quantifier is encountered; Urbach and Kutas (2010) and Nieuwland and colleagues (2010) showed rapid effects of quantification downstream of a quantifier, but those studies were not designed to test effects at the quantifier itself.

Since the violation in this condition relies on computing the pragmatic reading of *some* (that is to say, "not all"), the presence of an N400 effect suggests that this pragmatic reading is computed without any substantial delay. Such a finding is consistent with findings from several visual world eye-tracking paradigms suggesting that listeners become aware of the scalar implicature as soon as *some* is heard (Tavano, 2010; Grodner et al., 2010; Degen & Tanenhaus, 2010; but see Huang et al., 2010; Huang & Snedeker 2009, for alternative findings). The finding that the pragmatic reading of *some* becomes available just as quickly as the inherent logical reading is consistent with a default model of scalar implicature generation (see Introduction)—as

the present study does not include a condition to directly test when the logical reading of *some* becomes available it is not possible to make a direct comparison between when the pragmatic an logical meanings are computed, but the fact that the pragmatic violation elicited effects in the 150-300 ms time window, which is the earliest that lexical effects are ever consistently seen in violation paradigms, suggests that the pragmatic meaning was available immediately.

An alternate explanation for the finding of an N400-like effect in the present study is that it reflects the response to seeing an unexpected quantifier and thus has nothing to do with pragmatic computation: in other words, when participants see an "all"-type picture they make a forward prediction that they will see an "all"-type quantifier, and are thus surprised when they see *some*. While this is indeed a possibility, the fact that violation effects differed between pragmatic and logic violations (especially in the late time window, and possibly in the 300-500 ms time window; see Section 3.2.2) suggests that the responses obtained reflect more than just violations of prediction, since we would expect roughly identical patterns of effects if participants were merely matching pictures to quantifiers. In particular, the possibility that logic violations may not have elicited a robust N400 suggests that participants were actively trying to integrate quantification information, rather than just watching for expected quantifiers. As described in the Predictions (Section 1.6), after a "some"-type picture, the quantifier *all* is unexpected but can still be compatible with the picture depending on what is said later in the sentence; thus, if participants do not show a robust N400 for this condition, we can assume they are still actively trying to process the sentence, whereas if they were simply matching "all"-type pictures to the quantifier *all* then we would expect a robust N400 to the unpredicted quantifier. The lack of a robust P300 effect—which has been associated with, among other things, unexpected stimuli (see, e.g., Coulson et al., 1998)—also suggests that participants were not

simply strategically matching pictures and quantifiers.[15] Thus, I take these results to suggest that the pragmatic meaning of *some* did indeed become available to the parser immediately in the present study.

The present finding does not rule out context-driven models, however, because in this experiment the quantification information was highly relevant to the mapping between sentence and picture and to the completion of the acceptability judgment task. Therefore, context-driven models may also predict scalar implicatures to be computed immediately in this experimental context once the participant decides they are always relevant. Furthermore, an inherent limitation in applying the violation paradigm (which was used in the present study) to this research question is that it provides at best an upper bound for the time course of scalar implicature generation: we can only directly measure processing of a violation (specifically, processing an input that does not match the expected input), rather than the generation of the implicature itself. This is discussed in more detail in part 5 of this paper ("Limitations").

4.2.2 Differences between pragmatic and logical processing

Pragmatic and logic violations were found to elicit qualitatively different effects in the late time window: pragmatic violations elicited a significant right-lateralized negativity, whereas logic violations did not elicit a significant effect. (As described in section 3.2.2, pragmatic and logic violations may also have elicited different effects in the 300-500 ms time window, but this difference only showed up in exploratory t-tests, not in the omnibus ANOVA). This is important because, without different effects between logic and pragmatic violations, it would be difficult to rule out the possibility that effects obtained reflect only a mismatch with participants' forward

[15] I thank Mante S. Nieuwland for pointing out this argument.

predictions about highly predictable linguistic input (all the sentences in the stimulus set conformed to roughly the same structure, and all had a quantifier or classifier phrase as the second constituent of the sentence). Under such an interpretation, though, both logic and pragmatic violations would be expected to elicit similar ERP responses, given that both quantifiers ought to be equally unexpected in the violating contexts. The fact that they elicited qualitatively different responses suggests that the effects obtained in the present study, at least in the late window, are not merely a result of encountering an unexpected word, but in fact reflect differences between the processing of implicature-based violations and that of purely logic-based violations.

Again, this pattern of results is consistent with a default account of implicature generation, which assumes that the pragmatic reading of "some" is available just as soon as its inherent logical meaning (and, by extension, should be available just as soon as the logical meaning of "all", which is the only meaning that quantifier has); such an account would predict that both elicit similar responses in the early time window, but are differentiated in a later time window, since the meaning of "some" can be re-assessed if the context requires it, whereas the meaning of "all" cannot.

An alternative explanation for possible differences in ERP responses to pragmatic and logical errors is that these conditions also differ in their truth values at the point of the quantifier. For example, when given a picture in which some chefs are chopping cucumbers and some chopping onions, and then the sentence fragment "In the picture, all…", a participant may expect the sentence to be continued truthfully as "all the chefs are wearing hats" or "all the chefs have mustaches", without mentioning the difference between the two pictures. No such sentences were included in the stimulus set, and thus once the participants became accustomed to the

stimuli, the expectations for "all"-type and "some"-type pictures ought to be the same; presently there is not sufficient data to test this prediction (that would require comparing the first and second halves of each session, or comparing the first block of each session against the rest of the session), although this could be tested in the future. Furthermore, in future research the truth-value of the "all"-type sentences at the quantifier position could be ensured by using similar design but with sentences that indicate the activity before quantifying it (e.g., "This is a picture of chefs chopping things. All of them are chopping cucumbers.").

In any case, the question of whether the difference between effects for pragmatic and logic violations was due to different underlying processes or due to different truth-values at the quantifier position does not change the overall conclusion, which is that the fact that there were differences between the pragmatic and logic violations suggests that the pragmatic effect obtained was not simply a result of strategic matching. Therefore, the difference between effects is unlikely to be due to the possibility of a truthful continuation in the logic condition.

It is noteworthy that the logic violation in the present study did not elicit a sustained positivity like that elicited by the syntactically unlicensed "all"-like adverb in the study by Jiang and colleagues (2009). It is not surprising that the present study found different results, given that the linguistic manipulations between the two studies are actually quite different. Nevertheless, the results of the present study are not entirely consistent with Jiang and colleagues' (2009) interpretation of the sustained positivity as reflecting difficulty in linking the quantifying adverb with its antecedent. Regardless of differences between the linguistic phenomena in the two studies, both studies' violations should involve some sort of failure to link a quantifier or quantifying adverb with its antecedent, yet only Jiang and colleagues' (2009) violations triggered a sustained positivity. This may suggest that the sustained positivity elicited

in that study is somehow modular to syntactic processing, since the sentences tested involved syntactic ill-formedness whereas those in the present study did not. An alternate possibility, however, is that participants in the present study chose not to perform a costly linking operation, given that it was still possible that the sentence might turn out to be correct; under such an explanation, Jiang and colleagues' (2009) interpretation of the late positivity as reflecting general linking difficulties would still be tenable, with the caveat that the late positivity may only emerge when participants are forced to try to link incompatible elements, whereas the positivity may not emerge if they can postpone linking and wait to see if an alternative antecedent (in the case of the present study, some other object in the picture) becomes available for linking.

4.2.3 Sustained negativity

An alternative interpretation of the results obtained is that the negativity does not represent an N400 effect (or does not represent *only* an N400 effect), but rather represents a sustained negativity, given that there is a negativity in the data until at least 900 ms after the appearance of the stimulus, which is longer than the typical N400 time window. Sustained negativities in language tasks most often have an anterior distribution (Pijnacker et al., 2011), but recent studies have also found sustained negativities with a similar topography as the N400 (Jiang et al., 2009; Pijnacker et al., 2011). Both these studies included pure lexico-semantic anomalies to elicit standard N400s for comparison with the N400 effect. The present study did not include such a condition—while such violations were present in the fillers, they were not systematically controlled to allow for ERP analysis. This makes it difficult to classify

the effect obtained as an N400 or a sustained negativity.[16] It is possible, of course, that the epoch analyzed in this study involved both an N400 (in the 300-500 ms time window) and a sustained negativity (in the 600-900 ms time window) and that these are independent components reflecting different processes.

What would a sustained negativity in the present results indicate? Pijnacker and colleagues (2011) found a sustained negativity elicited in response to statements that were inconsistent with a previously-assumed inference, and interpreted that negativity as having to do with "revising" an inference "to incorporate an exception" (479) or otherwise negotiating between two conflicting information sources. The results of the present study are not entirely consistent with the inference-cancellation or inference-revision account. While underinformative sentences may trigger similar inference-revision processes, logically false sentences should not, since the meaning of "all" is semantically constrained and cannot be revised (that is to say, it is not defeasible) like the meaning of "some". If a sustained negativity reflected inference revision, the underinformative sentences in the present study would be expected to elicit a sustained negativity and the logically false sentences would not. This is indeed the pattern of results that was found, although ERPs elicited by logic violations did show a non-significant trend towards negativity; it remains to be seen whether a significant effect emerges when more data are available. Given that the point at which participants are able to realize the violation may differ between the "all" and "some" sentences (see discussion in the previous section), it is difficult to make a controlled comparison between these two effects anyway.

[16] A follow-up study currently under way does include such a condition, and will allow for a direct comparison between the lexico-semantic N400 and the quantifier N400.

Jiang and colleagues (2009), found a centro-parietal sustained negativity on words following a semantically unlicensed "all"-like quantifying adverb (in an acceptability judgment task) or left-frontal sustained negativity in the 600-900 ms time window following the quantifier itself (in a passive reading task). They interpret this as reflecting second-pass reanalysis mechanisms (such as ignoring the adverb or replacing the noun it refers to). While the late negativity in the present study shows a different topography than the one found in Jiang and colleagues' (2009) passive-reading task and a different latency than that found in their acceptability judgment task, the results of the present study are nevertheless consistent with Jiang and colleagues' (2009) interpretation, especially if the logic violation does turn out to elicit a reliable negativity.

In the present design it was not possible to compare the negativity elicited by quantifiers with a pure lexico-semantic negativity, nor was it possible to look at ERPs to following words as Jiang and colleagues did (since the words following the quantifier were not controlled across conditions). Attempting to dissociate the N400 effect from the sustained negativity effect would be a fruitful line of future research. Furthermore, being able to dissociate late activity associated with quantification problems in general from that associated with only pragmatic violations (i.e., to dissociate general second-pass reanalysis from processing specifically related to inference cancellation or implicature cancellation) would offer a promising means to compare predictions made by competing psycho-pragmatic theories of implicature generation (for reviews see Katsos & Cummins, 2010; Noveck & Reboul, 2008; Noveck & Sperber, 2007).

4.2.4 On the elicitation of N400 effects

Repetition is known to reduce N400 amplitudes and lead to relatively flat waveforms in the N400 time window (Kutas & Federmeier, 2000), and for this reason repetition of critical words is generally avoided in linguistic experiments, particularly those looking to analyze N400 effects. The present study, however, provides converging evidence with several previous studies (Renoult & Debruille, 2011; Shang & Debruille, 2011) showing that across-condition differences on the N400 can still be elicited even when critical words are repeated frequently within an experiment—in this case, the critical words *yǒu de* (有的), "some", and *suǒyǒu de* (所有的), "all", each appeared 80 times in the stimulus set. Furthermore, in addition to being repeated within the experimental context, these are closed-class words that have a high frequency in real life, which should also reduce N400 amplitudes across the board; nevertheless, in the present study, the N400 component for these words was still subject to modulation by pragmatic context, consistent with previous results showing that closed-class words still show N400 effects for frequency (Münte et al., 2001) and semantic constraint (Van Petten & Kutas, 1991).

Most importantly, the present study demonstrated that the N400 is sensitive to pragmatic context regardless of lexico-semantic manipulation. This finding is consistent with the already extensive literature on discourse influences on the N400 (for a review see, e.g., Van Berkum, 2009).

5 Limitations

This study adopted a novel design to examine scalar implicature, while addressing several conceptual and methodological limitations of previous studies. The study, however, has some of its own methodological issues that bear mention; some of these are being addressed in follow-up studies that are currently ongoing, whereas others are inherent to the design itself. The purpose of this section is both to outline how these limitations ought to be taken into account when interpreting the results of the present study, and hopefully to provide design guidelines for future ERP research using picture-sentence designs to investigate implicature processing.

As mentioned above, much of the data recorded in the present study showed substantial ocular artifact, which is likely to be due to the stimulus design and the demands placed on participants. Participants were required to attend to large pictures with white backgrounds presented on the screen in an otherwise dark room, and attend carefully enough to remember many small details that may have been probed in the comprehension questions later. After the presentation of the picture, many participants found it difficult to refrain from blinking during the RSVP sentence presentation, which lasted about 5-6 seconds. Participants were told that they could blink while the picture was being presented, but many did not for the reasons described above: as it turns out, extracting enough information from the picture to complete the task required substantial attention on the participants' part. The pictures used in this study were semantically rich compared to those used in most previous picture studies (Tavano, 2010; Grodner et al., 2010; Degen & Tanenhaus, 2010b; Huang & Snedeker, 2009; Foppolo, 2007; Hunt et al., in prep.; but see the pictures in Wu & Tan, 2009), in that they depicted realistic scenes with a large number of details, whereas many previous studies used carefully controlled and highly stylized pictures (for instance, in Grodner et al., 2010, and Huang & Snedeker, 2009,

"having" an object was depicted by placing the object somewhere near the agent in perceptual space). The details that were of relevance to the participants' task were not always readily apparent based on simple rules like their location on the screen, meaning that participants had to attend rather carefully to them to find the information they needed to complete the task. Thus, the difficulty of the task and general discomfort of looking at bright pictures and slowly-presented sentences led to a large amount of artifact, forcing me to perform my preliminary analysis on only a small subset of participants. Other ERP studies using picture-sentence verification designs in English (Knoeferle et al., 2011; Knoeferle & Kutas, 2011; Hunt et al., in prep.) have used a much faster stimulus presentation rate, with stimulus-onset asynchronies of about 500 ms and sentences that take about 2.5 to 3 seconds to present, which is about twice as fast as the sentences in the present design; these studies have also not experienced as much ocular artifact as the present study.[17] These studies also used a task that was probably less demanding (acceptability judgments), and gave participants more time to inspect the picture—Knoeferle and colleagues (2011) allowed participants to inspect the picture for as long as they want, while Hunt and colleagues (in prep.) used a fixed duration but one that was substantially longer than that used in the present study. I am now replicating this study with a more comfortable stimulus presentation paradigm (faster sentence presentation using auditory stimuli, longer picture presentation time, longer inter-trial intervals, more frequent breaks, and recording divided into two sessions) in attempt to collect data with less artifact.

Furthermore, in the present design, participants were not given explicit instructions about the nature of the acceptability judgment task, and thus may have been unsure what standards were expected of them when they made judgments. Instructions were also not carefully

[17] Pia Knoeferle, personal communication; Lamar Hunt, personal communication.

controlled across participants; while all participants were presented with the same instructions on-screen regarding the basics of the experimental procedure, I provided additional guidance encouraging participants to respond based on their intuitions only in cases when the participant requested clarification; therefore, the amount of detail in the instructions differed across participants. It has been shown that the wording of the instructions has a major effect on how participants evaluate and process scalar implicature (e.g., Bott & Noveck, 2004). In currently underway and future follow-up studies, participants are all given the same written instruction sheet before the experiment, and experimenters' verbal explanation of the task is kept to a minimum.

Another potential concern is the high predictability of "some"- and "all"-type quantifiers in the stimulus set used in the present study. Previous research has suggested (although it has not yet been confirmed in a direct comparison) that the judgment of scalar implicatures may be affected by the types of alternatives that are available; that is to say, participants may process scalar implicatures differently in an experimental context where "some" and "all" are their only options, than in a context that allows a more fine-grained scale, particularly one with numbers (Grodner et al., 2010; Degen & Tanenhaus, 2010a; Huang et al., 2010). Unfortunately, the ERP methodology requires a large number of items, and increasing the stimulus set to allow for a more enriched set of lexical alternatives would have been prohibitive in the current design; it may be possible to do so in future studies with a smaller set of critical stimuli (for instance, studies limiting themselves to pragmatic violations, rather than comparing pragmatic and logic violations). The present study attempted to minimize this issue by including other quantifiers in the fillers, but even so, most of the quantifiers were still similar in meaning to *yǒu de* (有的), "some", and *suǒyǒu de* (所有的), "all". Furthermore, including other quantifier types in the

fillers introduces another concern: if the proportion of fillers followed by certain kinds of violations (e.g., objects or verbs that do not match those portrayed in the pictures) differs from the proportion of critical sentences followed by the same violations, and fillers have different quantifiers (or numerals) than critical sentences, then the filler-specific quantifiers or numerals may act as a signal for increased likelihood of some type of violation downstream and the critical-trial-specific quantifiers may act as a signal for reduced likelihood of this type of violation, and thus effects that emerge on the critical quantifiers may represent differential predictive processing rather than scalar implicature processing. Therefore, the proportion of violation types must be carefully balanced across quantifier types, which clearly is likely to increase the number of stimuli required in an experiment. It remains to be seen whether these issues will be prohibitive to future ERP research on scalar implicature using this design.

As described above (section 4.2.3, "N400 versus sustained negativity"), a limitation of the present study is the inability to compare the quantifier-related (both pragmatic and logic) violations against a typical lexico-semantic N400 within the same experiment. A follow-up study that is currently under way (see Appendix I) includes a carefully-controlled object mismatch condition, in which the object named in the sentence does not match any of the items shown in the picture; this condition should be conducive to such a comparison.

Finally, one of the main aims of the present study was to dissociate the effects of *generating* a scalar implicature from the effects of *expecting* a content word based on an already-generated implicature. This is the reason the picture-sentence verification design was adopted, to make the infelicitousness of the critical sentences noticeable at the position of the quantifier itself. The idea is that, for the violation to be realized, the participant must first activate the pragmatic meaning of the quantifier by processing the implicature; thus, at the very least, the latency of the

effect should provide an upper bound for the time within which the implicature was processed. Furthermore, the differential responses to logic and pragmatic violations provided evidence that the results obtained reflect more than just expectation. The violation paradigm does, however, remain a conceptual limitation of this and previous studies on scalar implicature: the results recorded in such studies demonstrate brain activity associated with *violating* an expectation that is based on an implicature (and coping with that violation), rather than the effects of *generating* the implicature itself. That is to say, none of the responses obtained in the present study demonstrate the effect of generating a scalar implicature; we still do not know what sort of brain activity is elicited by the generation of an implicature. Furthermore, as pointed out by Tavano (2010), in scalar implicature research, averaging across trials in a violation paradigm may yield results that do not accurately reflect individual trials, as it is possible that participants only process the scalar implicature naturally in the first one or two trials (i.e., when they first need to decide how to respond to underinformative sentences), and their responses on subsequent trials merely reflect the output of a conscious decision-making process, rather than the online processing of the implicature. For all these reasons, caution must be exercised in implementing a violation paradigm to study scalar implicature generation.

A more straightforward way to measure brain activity associated with generating an implicature would be to adopt a design similar to those used by Breheny and colleagues (2006, Experiment 3; 2010) and Bergen and Grodner (2010), in which the experimenter does not create violations, but rather manipulates whether or not a scalar implicature is relevant given the context. For instance, Breheny and colleagues (2006, Experiment 3) manipulated whether the context was upper-bounding and therefore the implicature relevant (as in 11a), or the context lower-bounding and the implicature irrelevant (as in 11b):

11a) Mary asked John whether he intended to host all of his relatives in his tiny apartment. John replied that he intended to host **some of his relatives**. (The rest would stay in a hotel.)

11b) Mary asked John why he was cleaning his apartment. John replied that he intended to host **some of his relatives**. (The rest would stay in a hotel.)

Such a design would be a straightforward way to probe for ERP effects associated with the generation of scalar implicatures; context-based theories of scalar implicature generation tend to propose that an implicature will be generated in (11a) and not (11b), whereas default-based theories propose that an implicature will be generated in both sentences and will additionally be cancelled in (11b) (Katsos & Cummins, 2010). Either way, the comparison between the two conditions yields a straightforward subtraction. This sort of design would also probably require fewer stimuli and fillers to make the relevant comparison (since there is no need to distract the subject from pragmatic violations by introducing other violations in the fillers), thus perhaps leaving more room to manipulate the alternative quantifiers that are present (see the discussion above for why this may be important). Similar designs have been used in later studies by Bergen and Grodner (2010) and Breheny and colleagues (2010), which varied whether or not the speaker can be assumed to be knowledgeable (since scalar implicature generation, according to Gricean pragmatics, relies on the assumption that the speaker fully understands the context and consciously chose not to use a term higher on a scale). These are worthwhile directions in which to extend the current study.

6 Conclusion

The present study adapted a picture-sentence verification design to investigate the time course and electrophysiological instantiation of scalar implicature processing. It is the first electrophysiological study on scalar implicature to be able to measure effects the moment a scalar term is encountered, and the first online study on scalar implicature to investigate a non-Indo-European language. Quantifiers that were pragmatically incongruous with a context were compared with those that were logically incongruous with the context; it was found that pragmatically underinformative quantifiers elicited an early N400 as well as a late, right-lateralized negativity, and that the response to pragmatically underinformative quantifiers differed from that to logically incorrect quantifiers at least in the late time window. This suggests, first of all, that scalar implicatures are generated at an immediate level and information provided by a scalar implicature is processed as quickly as inherent semantic information; secondly, revision-related processes are carried out after it is realized that the automatically-generated pragmatic reading of the quantifier conflicts with the context. These results are easily accommodated by a default model of scalar implicature generation, although they do not necessarily rule out context-based models. Finally, this paper has presented several avenues for further study to extend and more deeply understand the results discussed here.

7 References

Baggio, G.; van Lambalgen, M.; Hagoort, P. (2008). "Computing and recomputing discouse models: an ERP study". *Journal of Memory and Language* 59, 36-53.

Bergen, Leon; Grodner, Daniel (2010). "Scalar implicatures are sensitive to the speaker's epistemic state". *23rd CUNY Conference on Human Sentence Processing.*

Bonin, Patrick; Peereman, Ronald; Malardier, Nathalie; Méot, Alain; Chalard, Maryl ène (2003). "A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies". *Behavior Research Methods, Instruments, & Computers* 35(1), 158-67.

Bornkessel-Schlesewsky, Ina; Schlesewsky, Matthias (2009). *Processing Syntax and Morphology: A Neurocognitive Perspective*. Oxford Surveys in Syntax and Morphology, 6. Oxford: Oxford University Press.

Bott, Lewis; Noveck, Ira (2004). "Some utterances are underinformative: the onset and time course of scalar inferences". *Journal of Memory and Language* 51, 437-57.

Breheny, Richard; Ferguson, Heather Jane; Katsos, Napoleon (2010). "Taking the epistemic step". *23rd CUNY Conference on Human Sentence Processing.*

Breheny, Richard; Katsos, Napoleon; Williams, John (2006). "Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences". *Cognition* 100, 434-63.

Chi, Weidong (2000). "Towards a logical differentiation between 'part' and 'some'". *Shandong Normal University Journal (Social Science)* 169, 91-103.

Coulson, Seanna; King, Jonathan; Kutas, Marta (1998). "Expect the unexpected: event-related brain response to morphosyntactic violations". *Language and Cognitive Processes* 13(1), 21-58.

Coulson, Seanna (2007). "Electrifying results: ERP data and cognitive linguistics". In M. Gonzalez Marquez, I. Mittelberg, S. Coulson, and M. Spivey (eds.). *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins.

De Neys, Wim; Schaeken, Walter (2007). "When people are more logical under cognitive load: Dual task impact on scalar implicature". *Experimental Psychology* 54(2), 128-33.

Degen, Judith; Tanenhaus, Michael (2010a). "Naturalness of lexical alternatives influences interpretation of 'some'". *23rd CUNY Conference on Human Sentence Processing.*

Degen, Judith; Tanenhaus, Michael (2010b). "When contrast is salient, pragmatic 'some' precedes logical 'some'". *23rd CUNY Conference on Human Sentence Processing.*

Degen, Judith (2009). *Processing Scalar Implicatures: An Eye-Tracking Study*. M.Sc. thesis, University of Osnabrück.

Fang, Ruifen (2008). "Reflections on the researches into scalar implicature". *Journal of Anhui Normal University (Humanities and Social Sciences)* 36(5), 594-8.

Fischler, Ira; Bloom, Paul; Childers, Donald; Roucos, Salim; Perry, Nathan (1983). "Brain potentials related to stages of sentence verification". *Psychophysiology* 20(4), 400-9.

Foppolo, Francesca (2007). "Between 'cost' and 'default': a new approach to scalar implicature. In R. Artstein & L. Vieu (eds.), *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, 125-31.

Grice, H.P. (1989). *Studies in the Way of Words*. Cambridge: Harvard University Press.

Grice, H.P. (1975). "Logic and conversation". In P. Cole and J. Morgan (eds.), *Syntax and semantics 3: Speech acts*. New York: Academic press. 41-58.

Grodner, Daniel; Klein, Natalie; Carbary, Kathleen; Tanenhaus, Michael (2010). "'Some,' and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment". *Cognition* 116, 42-55.

Hagoort, Peter; Wassenaar, Marlies; Brown, Colin (2003). "Syntax-related ERP effects in Dutch". *Cognitive Brain Research* 16, 38-50.

Horn, Laurence (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles.

Huang, Yi Ting; Hahn, Noemi; Snedeker, Jesse (2010). "Some inferences still take time: prosody, predictability, and the speed of scalar implicatures". *23rd CUNY Conference on Human Sentence Processing.*

Huang, Yi Ting; Snedeker, Jesse (2009). "Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface". *Cognitive Psychology* 58, 376-415.

Hunt, Lamar; Minai, Utako; Fiorentino, Robert; Politzer-Ahles, Stephen (in preparation). *An ERP study characterizing the time course of scalar implicature computation*. Undergraduate honors thesis, University of Kansas.

Jiang, Xiaoming; Tan, Yingying; Zhou, Xiaolin (2009). "Processing the universal quantifier during sentence comprehension: ERP evidence". *Neuropsychologia* 47, 1799-815.

Jiang, Xiaoming; Zhou, Xiaolin (2009). "Processing different levels of syntactic hierarchy in sentence comprehension: An ERP study on Chinese". *Neuropsychologia* 47, 1282-93.

Katsos, Napoleon; Cummins, Chris (2010). "Pragmatics: from theory to experiment and back again". *Language and Linguistics Compass* 4/5, 282-95.

Kliegl, Reihold; Wei, Ping; Dambacher, Michael; Yan, Ming; Zhou, Xiaolin (2011). "Experimental effects and individual differences in linear mixed models: estimating the

relationship between spatial, object, and attraction effects in visual attention". *Frontiers in Psychology: Quantitative Psychology and Measurement* 1, 1-12.

Knoeferle, Pia; Urbach, Thomas; Kutas, Marta (2011). "Comprehending how visual context influences incremental sentence processing: insights from ERPs and picture-sentence verification". *Psychophysiology* 48, 495-506.

Knoeferle, Pia; Kutas, Marta (2011). "Picture-sentence verification in older adults: evidence from ERPs". *18th Annual Meeting of the Cognitive Neuroscience Society*.

Kutas, Marta; Federmeier, Kara (2000). "Electrophysiology reveals semantic memory use in language comprehension". *Trends in Cognitive Sciences* 4(12), 463-70.

Kutas, Marta; Hillyard, Steven (1984). "Brain potentials during reading reflect word expectancy and semantic association". *Nature* 307, 161-3.

Landauer, T.K.; Foltz, P.W.; Dumais, S.T. (1998). "Introduction to latent semantic analysis". *Discourse Processes* 25, 259-84.

Lau, Ellen; Phillips, Colin; Poeppel, David (2008). "A cortical network for semantics: (de)constructing the N400". *Nature Reviews Neuroscience* 9, 920-33.

Li, Charles; Thompson, Sandra (1981). *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

Li, X.; Zhou, Xiaolin (2010). "Who is *ziji*? ERP responses to the Chinese reflexive pronoun during sentence processing". *Brain Research* 1331, 96-104.

Luck, Steven (2004). *An Introduction to the Event-Related Potential Technique*. Cambridge: MIT Press.

Münte, Thomas; Wieringa, Bernardina; Weyerts, Helga; Szentkuti, Andras; Matzke, Mike; Johannes, Sönke (2001). "Differences in brain potentials to open and closed class words: class and frequency effects". *Neuropsychologia* 39, 91-102.

Nieuwland, Mante; Ditman, Tali; Kuperberg, Gina (2010). "On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities". *Journal of Memory and Language* 63, 324-46.

Nieuwland, Mante; Kuperberg, Gina (2008). "When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation." *Psychological Science* 19(12), 1213-8

Noveck, Ira; Reboul, Anne (2008). "Experimental pragmatics: A Gricean turn in the study of language". *Trends in Cognitive Science* 12(11), 425-31.

Noveck, Ira; Sperber, Dan (2007). "The why and how of experimental pragmatics: the case of scalar inferences". In *Advances in Pragmatics*.

Noveck, Ira; Posada, Andres (2003). "Characterizing the time course of an implicature: an evoked potentials study". *Brain Research* 85, 203-10.

Oldfield, R.C. (1971). "The assessment and analysis of handedness: The Edinburgh inventory". *Neuropsychologia* 9, 97-113.

Rullman, Hotze; You, Aili (2006). "General number and the semantics and pragmatics of indefinite bare nouns in Mandarin Chinese". In Klaus von Heusinger and Ken P. Turner (eds.), *Where Semantics Meets Pragmatics*. Amsterdam: Elsevier. 175-96.

Pijnacker, Judith; Geurts, Bart; van Lambalgen, Michiel; Buitelaar, Jan; Hagoort, Peter (2011). "Reasoning with exceptions: an event-related brain potentials study". *Journal of Cognitive Neuroscience* 23(2), 471-80.

Renoult, Louis; Debruille, J. Bruno (2011). "N400-like potentials and reaction times index semantic relations between highly repeated individual words". *Journal of Cognitive Neuroscience* 23(4), 905-922.

Shang, Miles; Debruille, J. Bruno (2011). "N400 as an index of inhibition: a study with highly repeated stimuli". *18th Annual Meeting of the Cognitive Neuroscience Society.*

Szekely, Anna; Jacobsen, Thomas; D'Amico, Simona; Devescovi, Antonella, et al. (2004). "A new on-line resource for psycholinguistic studies". *Journal of Memory and Language* 51(2), 247-50.

Tavano, Erin (2010). *The Balance of Scalar Implicature.* Ph.D. thesis, University of Southern California.

Tsai, Wei-Tien Dylan (2004). "On *yǒu rén, yǒu de rén,* and *yǒu xiē rén*". *Chinese Linguistics* 8(2), 16-25.

Tsai, Wei-Tien Dylan (2003). "Three types of existential quantification in Chinese". In Audrey Li and Andrew Simpson (eds.), *Functional Structure(s), Form and Interpretation: Perspectives from Asian Languages*. London: RoutledgeCurzon.

Urbach, Thomas; Kutas, Marta (2010). "Quantifiers qualify more or less online: ERP evidence for partial incremental interpretation". *Journal of Memory and Language* 63, 158-79.

Van Berkum, J. J. A. (2009). "The neuropragmatics of 'simple' utterance comprehension: An ERP review". In U. Sauerland, & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory*. Basingstoke: Palgrave Macmillan

Van Petten, Cyma; Kutas, Marta (1991). "Influences of semantic and syntactic context on open- and closed-class words". *Memory and Cognition* 19, 95-112.

Wu, Zhuang; Tan, Juan (2009). "Scalar implicature in Chinese child language: An experimental study". *Journal of Foreign Languages* 32(3), 69-75.

Xie, Ying (2003). "On the construction of '*Yǒu de*（有的）+VP'". *Studies in Language and Linguistics* 23(3), 37-42.

Ye, Zheng; Zhou, Xiaolin (2008). "Involvement of cognitive control in sentence comprehension: Evidence from ERPs". *Brain Research* 1203, 103-115.

Ye, Zheng; Zhan, Weidong; Zhou, Xiaolin (2007). "The semantic processing of syntactic structure in sentence comprehension: An ERP study". *Brain Research* 1142, 135-45.

Ye, Zheng; Luo, Yue-Jia; Friederici, Angela; Zhou, Xiaolin (2006). Semantic and syntactic processing in Chinese sentence comprehension: Evidence from event-related potentials. *Brain Research* 1071, 186-96.

Yu, Jing; Zhang, Yaxu (2008). "When Chinese semantics meets failed syntax". *NeuroReport* 19(7), 745-9.

Zhang, Yaxu; Yu, Jing; Boland, Julie (2010). "Semantics does not need a processing license from syntax in reading Chinese". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36(3), 765-81.

Zhou, Xiaolin; Jiang, Xiaoming; Ye, Zheng; Zhang, Yaxu; Lou, Kaiyang; Zhan, Weidong (2010). "Semantic integration processes at different levels of syntactic hierarchy during sentence comprehension: An ERP study". *Neuropsychologia* 48, 1551-62.

Zhu, Xiaomei; Wang, Dandan (2007). "Syntactic effect on the generation of scalar implicatures". *Shandong Foreign Language Teaching Journal* 116, 68-71.

Appendix I. Rating study

A follow-up to this study, also using the ERP method, used a rating task (Urbach & Kutas, 2010, Experiments 2 & 3; Degen & Tanenhaus, 2010a; Foppolo, 2007), which allowed us to collect graded acceptability judgments about underinformative sentences and compare them with those for other types of violations. At the time of this writing, data from 16 participants has been collected.

Stimuli followed roughly the same design as that of the present study, and included the following critical conditions: a) 40 correct sentences; b) 40 sentences with correct objects but an underinformative quantifier "some"; c) 40 sentences with a correct quantifier but an object that was not present in the picture; and d) 40 sentences with both an underinformative quantifier and an incorrect object (referred to as a "double violation"). The stimuli underwent a sentence completion test to ensure that the correct objects used in the sentences were consistent with those participants expected based on the picture; all objects had a cloze probability of 46% or higher (average 81%), whereas incorrect objects all had a cloze probability of 0%. Critical sentences written so that none of the critical objects were at the end of the sentence. An additional 240 fillers of the following types were included:

- 120 correct "all" sentences
- 40 logically incorrect "all" sentences
- 80 sentences using other quantifiers:
    - 20 correct "some"-like
    - 20 correct "all"-like
    - 10 "some"-like with objects that did not match those shown in the picture
    - 10 "all"-like with objects that did not match those shown in the picture
    - 10 "some"-like with verbs that did not match the activity in the picture
    - 10 "all"-like with verbs that did not match the activity in the picture.

No fillers or critical sentences included grammatical or semantic anomalies (the sentences themselves were acceptable, but inconsistent with the pictures). The overall proportion of

inconsistent sentences in the experiment was 50% (assuming that pragmatic violations were treated as errors).

The stimulus presentation procedure was roughly the same as that of the present study, except that recorded stimuli were presented auditorily. After the spoken sentence finished playing, the screen displayed a scale from 1 to 7, with one side labeled "一致" ("consistent") and the other labeled "不一致" ("inconsistent"). The sides of the scale on which each label appears, and thus the hands used for adjusting judgments, were counterbalanced across participants. Participants were instructed to rate each sentence based on how well it described the characters, objects, and events portrayed in the picture (the instructions did not mention numbers of characters or objects); they were all instructed that "there is no right or wrong answer", and that "the experiment's goal is to measure your own language intuitions". They submitted their responses with the right and left buttons of a gamepad. They had 3000 ms to submit a response; the rating scale disappeared when they submitted a response or when 3000 ms passed.

For the behavioral responses, data from participants whose scale was reversed were re-coded such that, across all participants, "1" represented "inconsistent" and "7" represented "consistent". The average rating for each condition is given in Table 4:

Condition	**Rating**
Correct "some"	6.319
Underinformative quantifier	5.494
Mismatching object	3.340
Double violation	2.454
Correct "all"	6.179
Logically false "all"	2.101

Table 4: Rating results for each sentence condition. 7 refers to "most consistent", 1 refers to "most inconsistent".

Pairwise t-tests revealed significant differences between all of these conditions except correct "some" and correct "all", and double violation and logically false "all". In short: both correct types were judged as the most consistent with the pictures, underinformative quantifier types judged as less consistent, mismatching object types less consistent still, and double violation (both mismatching object and underinformative quantifier) and logically false quantifier types the least consistent. Of greatest relevance to the current study is the finding that underinformative sentences, while judged worse than correct sentences, are judged better than both logically false and object-mismatching sentences (both of which were included in the present study as well).

Contrast	***t***	**df**	***p***
Correct "some" – Mismatch	10.361	15	<0.000
Correct "some" – Underinf.	5.020	15	<0.000
Correct "some" – Double	22.397	15	<0.000
Correct "some" – Correct "all"	1.785	15	0.095
Correct "some" – Logically false	20.286	15	<0.000
Mismatch – Underinf.	-5.813	15	<0.000
Mismatch – Double	3.987	15	<0.000
Mismatch – Correct "all"	-9.081	15	<0.000
Mismatch – Logically false	4.080	15	0.001
Underinf. – Double	12.580	15	<0.000
Underinf. – Correct "all"	-4.040	15	0.001
Underinf. – Logically false	11.668	15	<0.000
Double – Correct "all"	-20.530	15	<0.000
Double – Logically false	2.327	15	0.034
Correct "all" – Logically false	20.549	15	<0.000

Table 5: Pairwise *t*-tests for each combination of sentence conditions.